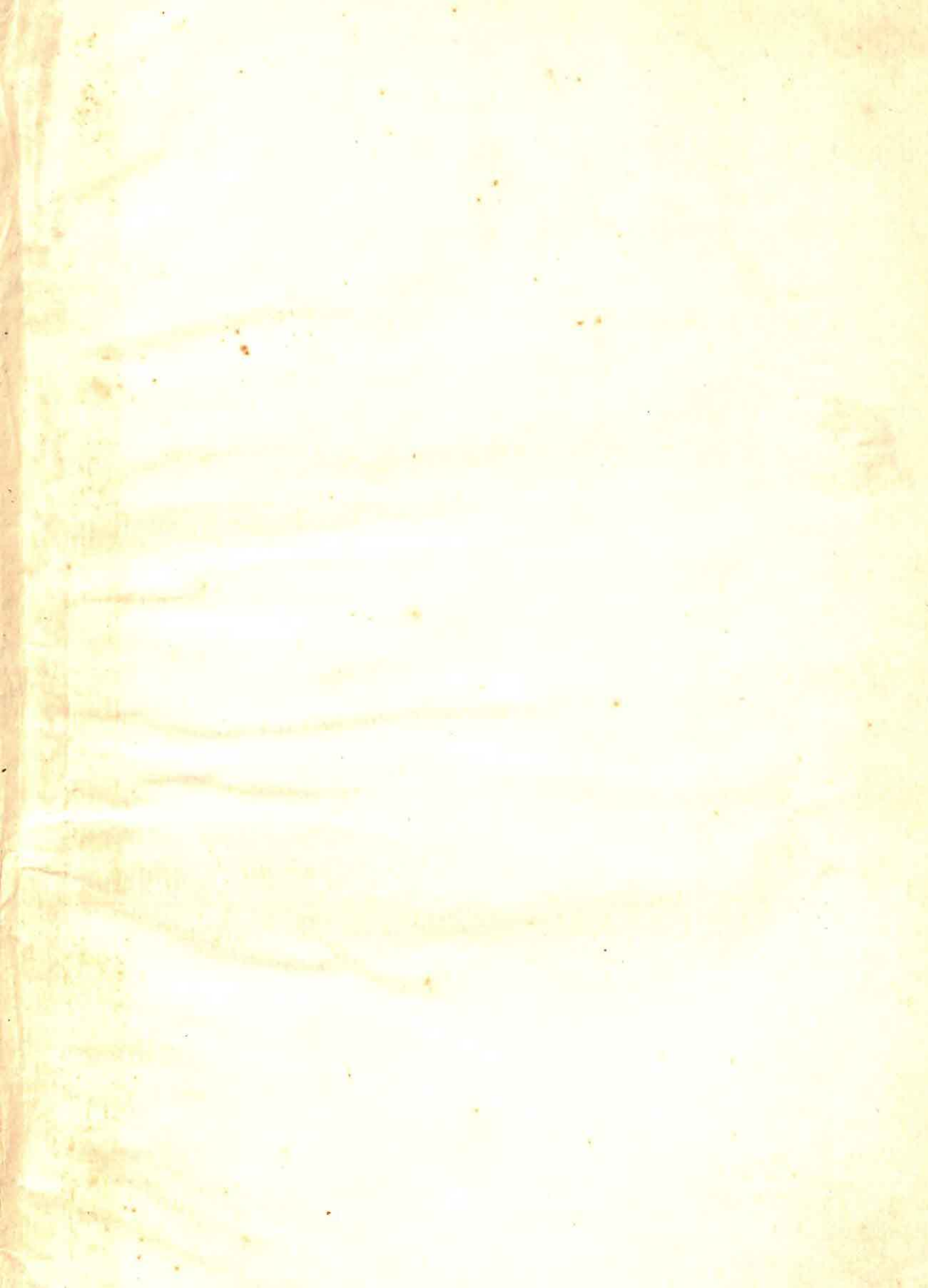


15/2



Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology*
with the *Psychological Monographs*

VOL. 71

1957

HERBERT S. CONRAD, Editor
Department of Health, Education, and Welfare
Office of Education
Washington 25, D.C.

Consulting Editors

DONALD E. BAIER
FRANK A. BEACH
ROBERT G. BERNREUTER
WILLIAM A. BROWNELL
HAROLD E. BURTT
JERRY W. CARTER, JR.
CLYDE H. COOMBS
JOHN F. DASHIELL
EUGENIA HANFMANN
EDNA HEIDBREDER

HAROLD E. JONES
DONALD W. MACKINNON
LORRIN A. RIGGS
CARL R. ROGERS
SAUL ROSENZWEIG
ROSS STAGNER
PERCIVAL M. SYMONDS
JOSEPH TIFFIN
LEDYARD R. TUCKER
JOSEPH ZUBIN

Published by

THE AMERICAN PSYCHOLOGICAL ASSOCIATION
1333 SIXTEENTH STREET N.W., WASHINGTON 6, D.C.



7. 9. 70
J - 366

CONTENTS OF VOLUME 71

Whole No.

- 430 USE OF THE SQUARE-ROOT METHOD TO IDENTIFY FACTORS IN THE JOB PERFORMANCE OF AIRCRAFT MECHANICS. Charles Wrigley, Charles N. Cherry, Marilyn C. Lee, and Louis L. McQuitty. Pp. 28.
- 431 THE EFFECT OF VARYING AMOUNTS AND KINDS OF INFORMATION AS GUIDANCE IN PROBLEM SOLVING. Bernard R. Corman. Pp. 21.
- 432 FAILURE-AVOIDANCE IN SITUATIONAL INTERPRETATION AND PROBLEM SOLVING. Harold M. Schroder and David E. Hunt. Pp. 22.
- 433 PURE-TONE THRESHOLDS FOLLOWING STIMULATION BY NARROW-BAND FILTERED NOISE. John L. Fletcher. Pp. 13.
- 434 EFFECTS OF LIGHT ON ELECTRICAL EXCITATION OF THE HUMAN EYE. Lorrin A. Riggs, Janet C. Cornsweet, and Warren G. Lewis. Pp. 45.
- 435 A FACTOR-ANALYTIC STUDY OF PLANNING ABILITIES. R. M. Berger, J. P. Guilford, and P. R. Christensen. Pp. 31.
- 436 INFANT DEVELOPMENT UNDER ENVIRONMENTAL HANDICAP. Wayne Dennis and Pergrouhi Najarian. Pp. 13.
- 437 GRADIENTS OF ERROR-REINFORCEMENT IN A SERIAL PERCEPTUAL-MOTOR TASK. Melvin H. Marx. Pp. 20.
- 438 THE IN-BASKET TEST. Norman Frederiksen, D. R. Saunders, and Barbara Wand. Pp. 28.
- 439 FORCED CHOICE AND OTHER METHODS FOR EVALUATING PROFESSIONAL HEALTH PERSONNEL. Sidney H. Newman, Margaret A. Howell, and Frank J. Harris. Pp. 27.
- 440 IMPULSE EXPRESSION AS A VARIABLE OF PERSONALITY. Nevitt Sanford, Harold Webster, and Mervin Freedman. Pp. 21.
- 441 A BEHAVIORAL CENSUS OF A STATE HOSPITAL POPULATION. Stanton P. Fjeld, and others. Pp. 31.
- 442 THE RETINAL SIZE OF A FAMILIAR OBJECT AS A DETERMINER OF APPARENT DISTANCE. Walter C. Gogel, Bryce O. Hartman, and George S. Harkker. Pp. 16.
- 443 THE EFFECT OF THREE TEACHING METHODS ON ACHIEVEMENT AND MOTIVATIONAL OUTCOMES IN A HOW-TO-STUDY COURSE. John D. Krumboltz and William W. Farquhar. Pp. 26.
- 444 A CRITERION FOR COUNSELING. Clifford P. Froehlich. Pp. 12.
- 445 AGGRESSION IN FANTASY AND OVERT BEHAVIOR. Arthur R. Jensen. Pp. 13.
- 446 A PATTERN ANALYSIS OF DESCRIPTIONS OF "BEST" AND "POOREST" MECHANICS COMPARED WITH FACTOR-ANALYTIC RESULTS. Louis L. McQuitty. Pp. 24.
- 447 CONTRIBUTIONS TO THE STUDY OF THE PROBLEM-SOLVING PROCESS. Etwin Roy John. Pp. 39.
- 448 SOME DIMENSIONS OF INTERPERSONAL RELATIONS IN THREE-MAN AIRPLANE CREWS. Benjamin Fruchter, Robert R. Blake, and Jane Srygley Mouton. Pp. 19.
- 449 AGGRESSION AND AGE IN RELATION TO VERBAL EXPRESSION IN NONDIRECTIVE PLAY THERAPY. Dell Lebo and Elaine Lebo. Pp. 12.
- 450 INDIVIDUAL DIFFERENCES IN WHOLE-PART APPROACH AND FLEXIBILITY-RIGIDITY IN PROBLEM SOLVING. Ralph H. Goldner. Pp. 18.

- 451 EFFECTS OF TASK MOTIVATION AND EXPECTANCY OF ACCOMPLISHMENT UPON ATTEMPTS TO LEAD. John K. Hemphill and others. Pp. 16.
- 452 THE PENSACOLA Z SURVEY: A STUDY IN THE MEASUREMENT OF AUTHORITARIAN TENDENCY. Marshall B. Jones. Pp. 19.
- 453 EFFECTS ON CLIENTS OF A REFLECTIVE AND A LEADING TYPE OF PSYCHOTHERAPY. Jellison D. Ashby, Donald H. Ford, Bernard G. Guerney, Jr., Louise F. Guerney, and William U. Snyder. Pp. 32.

Psychological Monographs: General and Applied

Use of the Square-Root Method to Identify Factors in the Job Performance of Aircraft Mechanics¹

CHARLES WRIGLEY,² CHARLES N. CHERRY,³
MARILYN C. LEE, AND LOUIS L. MCQUITT⁴
University of Illinois

I. INTRODUCTION

THIS report has a threefold purpose: (a) to exemplify application of a modified version⁵ of square-root factor analysis to a large correlation matrix, (b) to describe the factorial structure of

job performance ratings obtained from supervisors of aircraft mechanics,⁶ and (c) to compare the results of this study with those of two earlier and similar investigations (6, 7) in order to determine the most appropriate source of items for the rating of job performance in the field of aircraft maintenance.

Taking up the third purpose first, it should be explained that in the first study of the series (7), the descriptive items to rate job performance were obtained from interviews with supervisors of aircraft mechanics. In the second (6), items were selected from a wide variety of factor analyses of personality traits appearing in the psychological literature. In the third, reported here, items were drawn from existing rating scales and questionnaires designed to measure various aspects of job performance.

Each phase of the research led to preparation of a different inventory intended for use as a rating device. The development and analysis of each inventory was conducted in accordance with the same general plan. An inventory of 200 items was obtained by assembling

¹This research was supported in part by the United States Air Force under Contract No. 33-(038)-25726 monitored by the Air Force Personnel and Training Research Center. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

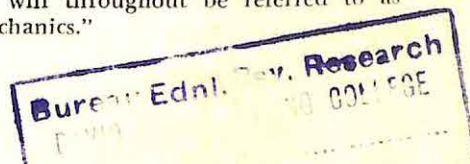
²Now at the University of California.

³Now at State Farm Insurance Companies, Bloomington, Illinois.

⁴This study was a cooperative project whose completion depended upon the assistance of many persons. Kern W. Dickman and Mrs. Carol R. Tucker carried out the statistical analyses. Mr. Dickman also prepared a computer program for rank-ordering sets of correlations and factor loadings which greatly reduced the amount of hand computation required. Ramona J. Russell and William L. Huffman supervised the computer operations, and Robert D. Boys and Marvin I. Knopp assembled and organized the statistical results. The authors wish to express their sincere appreciation to the persons named for their assistance and cooperation.

⁵The modified square-root method of factor analysis was developed by Wrigley and McQuitty (9) in an attempt to meet certain criticisms customarily advanced against the square-root method, with a view to making factor analysis more readily applicable to large correlation matrices. The nature and purposes of the changes are briefly described in Section III of this report.

⁶The subjects of this study were Air Force Aircraft and Engine Mechanics. To conserve space, they will throughout be referred to as "aircraft mechanics."



a pool of items and revising a sample of them. This was administered to a large sample of supervisors of aircraft mechanics. (N 's for the three studies ranged from 383 to 464.) Approximately one quarter of the subjects in each sample were asked to describe a "best," one half to describe an "average," and one quarter to describe a "poorest" mechanic, by responding "true" or "false" to the inventory items. Each inventory was then factorized in order to permit the development of a single, more comprehensive inventory in which overlapping of items in terms of factorial content would be minimal.

It should be noted here that while in principle factor analysis is generally considered to be the most satisfactory procedure currently available for reducing the number of items or variables with least loss of information, it suffers from a serious limitation, namely, computations are simply too laborious for application to any very large set of variables. In the present series of studies, where so many items had to be considered, a simpler procedure was needed than calculation and rotation of a centroid or a multiple-group factor analysis. The modified square-root method was accordingly used. A virtue of this procedure is that it can be applied without incommensurate labor to as many as 200 or 250 variables, provided reasonably good computational facilities are available. Because of its potential usefulness in other studies where the number of variables to be analyzed is very large, this report is devoted in part to a brief description of the method.

The organization of this monograph is as follows: The second section describes the preparation and administration of the job performance inventory, the third

describes the square-root method of factor analysis, and the fourth and fifth sections present the results of the statistical analyses of the data. In the sixth section, results of the present study are compared with those of the two which preceded it in the series. The study is summarized in the seventh and final section.

II. PREPARATION AND ADMINISTRATION OF THE "SURVEY INVENTORY"

The psychological literature contains many rating scales and questionnaires designed to measure various aspects of vocational proficiency. Few of these studies refer directly to the selection of mechanics. However, many deal with jobs which appear to be somewhat similar, and include items referring to traits of possible importance to success in aircraft maintenance work.

It was therefore hypothesized that items drawn from these studies would be useful for describing job performance of aircraft mechanics. The content of these scales is broad in scope, and each item reflects the accumulated experience of one or more experts in personnel psychology. Appropriate sampling of this material should thus contribute to the attempt to isolate characteristics related to efficient aircraft maintenance.

Collection and sampling of the item pool. In order to locate pertinent studies, the *Psychological Index* and *Psychological Abstracts* from 1900 to 1952 were consulted for articles containing relevant rating scales and questionnaires. The search yielded 64 studies containing descriptive phrases which appeared to be related to the on-the-job behavior of aircraft mechanics. From these sources, some 3,365 items were chosen.

These items were then classified into

42 content categories according to a method developed by Staugas and McQuitty (8), and a stratified sample was drawn from the pool in such a way as to minimize overlap with the two preceding inventories. In selecting items for the sample, preference was given to those which (a) were expressed in simple language and (b) appeared to represent most clearly the essence of the category into which they had been classified. To allow for rejection of items which might prove unsatisfactory upon preliminary administration, 221 rather than 200 items were initially selected from the pool.

Modification of items and assembly of the inventory. In assembling the inventory, the aim was to retain each item in the exact form in which it was written whenever possible, and if any alterations in wording had to be made, to insure that the original meaning of the item remained unchanged. The following types of editorial revisions were made:

1. Items calling for self-judgments were changed to a form requiring judgments of others. This was done because the inventory was intended for use by supervisors for the rating of line workers.

2. Items referring to supervisory personnel were modified for application to line mechanics.

3. Items relating to civilian jobs were reworded to apply more specifically to Air Force maintenance work.

4. Items designed for females were changed to apply to males.

5. Items in the form of a question were turned into declarative statements.

6. Some items were shortened to improve their clarity.

7. The grammatical form was rendered consistent from item to item.

8. Words judged by the investigators to be difficult were replaced by words believed to be more commensurate with the educational attainments of the average Air Force mechanic.

9. In a "double-barrelled" item, the more general phrase of the two was retained, and

the qualifying, subordinate, or more specific phrase eliminated.

10. All items were cast into the present or the conditional tense for consistency and ease of interpretation.

Upon completion of editing, the phrases were assembled into a preliminary inventory and administered individually to 28 aircraft mechanics, all of whom had had at least six months of experience on the line. Half of these subjects were asked to rate the "best" and half were requested to rate the "poorest" mechanic with whom they had worked, by checking "true" or "false" for each phrase in the inventory. (The two-point scale of "true" and "false" was decided upon because of our intention to use phi coefficients in the factor analysis.) The entire group of subjects was encouraged to make comments on the individual items.

After the responses had been tallied, chi square was calculated for each item in order to locate those which failed to discriminate significantly between the descriptions of "best" and "poorest" mechanics. The items were then reviewed by three judges. Those which met *all* of the following conditions were retained without modification:

1. The chi-square value was significant at the .05 level.

2. No comment concerning either wording or understanding of the item was offered during individual administration.

3. Neither answer alternative to the item was endorsed by fewer than 10 per cent of the subjects.

These requirements were met by 167 of the items. The remaining items were considered one at a time with respect to the above three criteria. Items with less than a 10 per cent response for either answer alternative were rewritten to enlarge this percentage whenever this appeared possible; otherwise they were re-

jected. Items which mechanics reported as requiring information that they did not have were rejected, and those difficult to understand were amended. This procedure yielded a final inventory of 200 items, named by us the Survey Inventory. This is the instrument used in all subsequent phases of this study. It is reproduced in the Appendix.

Administration of the inventory. To obtain data for the factor analysis, the inventory was administered to 464 supervisors of aircraft mechanics from Chanute, Sheppard, and Connally Air Force Bases. In this administration, 148 supervisors were asked to describe the "best," 179 were asked to describe an "average," and 137 were asked to describe the "poorest" mechanic with whom they had worked during the previous two years. The inventory was given in small group sessions (10 to 30 men) of about 30 minutes in length. This allowed ample time for all subjects to complete the items.

For each subject, the following occupational experience was required: (a) completion of the basic training course for aircraft mechanics; and (b) at least six months' experience as a supervisor of aircraft mechanics.

The data obtained in this phase of the study were then subjected to a modified square-root factor analysis. These results are presented in Section IV.

III. THE SQUARE-ROOT METHOD OF FACTOR ANALYSIS

The standard methods of factor analysis are generally agreed to be too laborious to be readily applicable to a correlation matrix much larger than 100 variables. Yet there are occasions when it would be desirable to analyze correlation matrices of a size considerably larger

than this. The square-root method of factor analysis (sometimes known as the diagonal method) is the simplest factor-analytic procedure, but has not won widespread favor because of various criticisms advanced against it. The method has accordingly been modified by Wrigley and McQuitty (9) in an attempt to meet these criticisms; in the course of doing so, the computations have been still further simplified.

Our restricted objective in making these modifications should be made quite explicit. The intention was not to provide a method which is in any sense to be regarded as a competitor of standard methods; the purpose was merely to provide some way of analyzing correlation matrices too large for the regular procedures. Hence it is not incumbent upon us to demonstrate that our solution is as satisfactory as, say, a simple structure solution would be, even were it possible to do so. It suffices if we can demonstrate that square-root factors are generally interpretable and psychologically meaningful, and add in some measure to our understanding of the structural interrelations of the variables analyzed.

The square-root method. First let us briefly summarize the standard form of the square-root method as well as the criticisms which have been made of it. The first step in the procedure is to calculate loadings for a first factor in such a way as to reduce all residual correlations with a selected *pivot variable* to zero. This is equivalent to placing a reference axis collinear with the test vector of the pivot variable. Residual correlations are then calculated in the usual way. This implies projection of the test vectors onto a plane orthogonal to the reference axis. In the next step, loadings are calculated for a second fac-

tor so that residual correlations for a second pivot variable are reduced to zero. The second axis will be collinear to the *residual* test vector of the second pivot variable, and will be at right angles to the first. The second set of residual correlations is calculated, and so the method proceeds. Since successive reference axes are at right angles, the factors are mutually orthogonal, so that the factor measurements for the persons in the sample, if calculated, would be uncorrelated.

The computational formula for square-root loadings is simple. The factor loading f_{i1} for the i th variable on the first factor is:

$$f_{i1} = \frac{r_{iq}}{\sqrt{r_{qq}}}, \quad [1]$$

where r_{iq} is the correlation of the i th variable with the pivot variable q and r_{qq} is the diagonal entry for the pivot variable.

More generally, the loading for the j th factor, with pivot variable s , is:

$$f_{ij} = \frac{r_{is}^{(j-1)}}{\sqrt{r_{ss}^{(j-1)}}}, \quad [2]$$

where $r_{is}^{(j-1)}$, $r_{ss}^{(j-1)}$ are the residual correlations for the $(j-1)$ th factor.

Criticisms of the method. There have been two main criticisms of the square-root method:

1. The procedure depends to a greater extent than most other factor-analytic methods upon accurate estimation of the communalities, since diagonal entries play such an important role in calculating the factor loadings. If some communalities are underestimated, then some diagonal entries may become negative during the calculations. This means that one must take the square root of a negative number, which results in factor loadings which are imaginary.

2. The order of selection of pivot variables is unspecified, so that p factorial solutions are possible (where p = the number of variables). No rule is provided for selecting any one order of pivot selection in preference to any other.

The modified procedure. Our modifications were intended to provide (a) a logical basis for specifying order of selection of pivot variables, (b) a method not dependent in any way upon the arbitrary judgment of the investigator, and (c) a computational procedure rapid enough to permit the handling of large correlation matrices. The following three modifications have accordingly been introduced:

1. Unities are inserted in the leading diagonal in place of communalities. This insures that so long as product-moment correlations are analyzed, no diagonal entry can become negative during the calculation.⁷ (Phi coefficients and point biserials will be remembered to be special forms of product-moment correlation.)

It is true that if unities are used, the factors obtained include specific as well as common variance. There are some who have doubted whether any factors derived by the use of unities rather than communalities can be psychologically meaningful. Our experience has been that they are interpretable. It will be noted that the specific variance in any factor is restricted to the loading for the pivot variable. The other loadings represent only common variance. (It can be shown that the sum of squares of loadings for any nonpivot variable is equal to the squared multiple correlation with the set of pivot variables, i.e., is a measure of common variance only.) An incidental result of using unities is that factor measurements, if required, can be calculated exactly instead of having to be estimated, as is necessary in any communalities solution (5, ch. 12).

2. The variable with the largest sum of off-diagonal correlations (and subsequent residual correlations) is selected as the pivot variable. This provides an objective criterion for pivoting order, and insures every factor will have some large loadings and that these will be predominantly positive.

3. A formula, based upon an algebraic identity, was developed for calculating the sum of residual correlations for each variable without having to compute each individual residual. In the extraction of each factor, therefore, there must be calculated only the p column-sums, followed by p residuals for the indicated pivot

⁷ Ford, L. R., Jr. A proof that the matrices of certain indices of association are Gramian. Unpublished manuscript, 1952.

variable, instead of the usual $p(p+1)/2$ residual correlations.

Nonrotation of square-root factors. Rotation is generally essential for a principal axes or a centroid solution because the axes tend to pass between, rather than through, the test clusters. In the square-root method, however, the reference axes are placed collinear with the pivot variables. Hence if each pivot lies in a fairly central position in a cluster, the reference axis will pass *through the cluster*. This generally turns out to be the case, since in the present method the pivot variables are selected on grounds of their high column-sums. For this reason, no rotation of the factor loadings is considered necessary.

Parenthetically, it should be noted that when the data are dichotomous, our preference is for calculating phi coefficients rather than tetrachoric correlations. The correlation matrix is then necessarily Gramian, so that no diagonal entry can become negative in the course of the analysis (unless as a result of rounding errors).

The amount of variance accounted for by any factor in an orthogonal solution is, as is well known, a function of the sum of squares of the loadings. In the principal axes method, the sum of squares of each successive factor is necessarily equal to or smaller than those for preceding factors, and this relation is generally true for the centroid method. In our method of analysis, factors are not necessarily extracted in order of contribution of variance. However, because of pivot selection depending on column sums, there is a marked tendency for larger factors to be extracted first and smaller factors later. (See Table 1 for evidence of this progression.) Our computational procedure has been to extract

more factors than needed, this being practicable because of the speed of the method, and then to order them according to sums of squares of loadings, retaining and reporting only the larger ones.

The generally smaller size of square-root loadings. The square-root loadings obtained by the above procedure tend, in general, to be lower than would be obtained by a standard solution, such as a simple structure rotation of a centroid analysis of tetrachoric correlations. There are three main reasons for this:

1. Phi coefficients are generally smaller than tetrachoric correlations.
2. Even for the same correlation matrix, square-root loadings tend to be lower than centroid loadings. The sum of squares of loadings for any nonpivot variable in a square-root solution can be shown to be equal to the squared multiple correlation (SMC) of that variable with the k pivot variables. But the SMC of any variable with the k pivot variables only is necessarily less than or equal to the SMC of the variable with the full set of $(p-1)$ other variables in the analysis. Guttman (4, pp. 92-93) has shown this latter to constitute a lower bound for the communality. Hence the sum of squares of loadings for any nonpivot variable is necessarily less than or equal to the communality. In practice it is generally appreciably less.
3. A general factor is more likely in a square root than in a simple structure solution, for the reason that the variable with largest column-sum is usually some rather generalized measure. Because more variance is accounted for by the first factor, less therefore remains to define later factors.

Any new psychometric procedure has to establish two points: (a) its computational feasibility, and (b) its psychological usefulness. The first requirement is clearly met in the present instance. The modified square-root method is considerably more rapid than any other factor-analytic procedure currently available, primarily because there is no longer any problem either of reflection or of rotation. In addition, the number of residual

correlations to be calculated in extracting each factor is much reduced. The computational gain is particularly evident on an electronic computer. On Illiac (the University of Illinois computer), approximately six factors can be extracted from a correlation matrix of order 200 in an hour. This time refers, of course, only to machine operation, and much further time may be needed for preparing tapes, printing results, etc. Even so, the factor analysis of a large correlation matrix becomes quite feasible.

The main problem is therefore whether the classification of variables provided by the square-root method is psychologically meaningful, although unities have been inserted in the diagonal, so that results may properly be used to reduce the number of variables to more manageable proportions. The data in the next section provide some of the evidence in support of our assertion that the factors provided by the modified square-root method are both interpretable and informative.

IV. RESULTS OF THE FACTOR ANALYSIS

In obtaining the phi coefficients for the factor analysis, all subjects (whether classified as "best," "average," or "poorest" mechanics) were included in a single sample. The square-root factor analysis of the phi coefficients was continued until 40 factors had been extracted. This extensive factoring was carried out to determine (a) the relative size of the successive factors, and (b) the number of square-root factors which appear to be psychologically meaningful, in a correlation matrix as large as this.

Table 1 presents sums of squares of loadings for the 40 factors and indicates their order of extraction. These factors

TABLE 1
SQUARE-ROOT FACTOR ANALYSIS: SUMS OF
SQUARES OF FACTOR LOADINGS

Factor in order of size	Factor in order of extraction	Sum of squares of factor loadings	Factor in order of size	Factor in order of extraction	Sum of squares of factor loadings
1	1	53.6714	21	17	1.3461
2	5	9.2086	22	16	1.3385
3	2	7.7215	23	18	1.3249
4	6	3.3006	24	27	1.2724
5	8	2.0307	25	25	1.2705
6	3	2.8628	26	28	1.2492
7	4	2.1757	27	21	1.1665
8	9	2.0308	28	23	1.1349
9	12	1.8699	29	38	1.1257
10	14	1.8280	30	36	1.0190
11	13	1.7899	31	26	.9507
12	29	1.7800	32	31	.9473
13	11	1.7171	33	32	.9434
14	7	1.6961	34	33	.9228
15	15	1.5993	35	35	.9169
16	22	1.5618	36	30	.9121
17	24	1.5545	37	34	.8230
18	19	1.5001	38	40	.8139
19	20	1.4529	39	39	.7894
20	10	1.4250	40	37	.7411

Factors	Sum of squares of factor loadings	Percentage contribution to total variance
1	53.6714	26.8%
2-10	33.9286	17.0
11-20	16.0767	8.0
21-30	12.2483	6.1
31-40	8.7606	4.4
Remaining variance	75.3144	37.7
Total (200 items)	200.0000	100.0%

together account for 62.3 per cent of the total variance. The total variance is based upon all items, and is equal to the number of items, since each item has been taken to have unit variance. If all factors had been extracted, the total sum of squares of loadings would have equalled the total variance (i.e., 200).

The 28 largest of the extracted factors appeared sufficiently meaningful to be named. However, owing to space limitations, only the first ten will be reported here.⁸ A second restriction involves the

⁸ A 19-page description of Factors 11 to 28 has been deposited with the American Documentation Institute. Order Document No. 4839 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D.C., remitting in advance \$1.75 for microfilm or \$2.50 for 25 photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

number of loadings presented for each factor. It is customary, in reporting factorial studies, to present all variables with loadings equal to or greater than a stated size—often .30—and to exclude all with loadings less than that. In this study, however, there were 173 loadings above .30 for the first factor, so that we have reported only the highest 30, since these suffice to describe the factor. With respect to the remaining factors, the conventional cutoff point of .30 would be inappropriate because, as previously pointed out, the square-root loadings for factors subsequent to the first are lower than their simple structure counterparts. We have therefore uniformly used the ten highest loadings, even when lower than .30, to determine whether the factor is meaningful.

In the following discussion we shall use the term *criterion correlation* to designate the phi coefficient between the "true" or "false" responses to each item and the supervisors' classifications of the mechanics as either "best" or "poorest," i.e., excluding "average" mechanics from the sample. The primary reason for ex-

cluding them in calculating these coefficients was the availability of an electronic computer program for dichotomous data. Exclusion of the average group was preferred to combining it with one of the other classes because we were interested only in the *comparative* indices of validity. By computing the criterion correlations on the extreme cases only, the interitem comparisons were made as sharp as possible.

It will no doubt be apparent that these criterion correlations are probably heavily loaded by halo effects. When the supervisor selects a "poorest" mechanic for description, he is obviously likely to describe him as using good judgment in most situations, or as being respected by his co-workers. However, the probable existence of such effects does not make the criterion correlations valueless. By means of the criterion correlations, we can determine which attributes form part of the supervisors' global impressions (or halo effects) and which do not.

The ten largest factors obtained by the square-root analysis will now be reported.

Factor 1. General Job Efficiency

Item number	Loading	Criterion correlation	Descriptive phrase
86	1.000	.775	Does jobs in order of importance
187	.764	.870	Fits well into the organization
116	.732	.865	Goes over own work to see that it was done correctly
161	.722	.816	Works well without supervision
131	.720	.809	Sees what will be needed later and gets ready for it
8	.693	.857	Seems to have a good future as a mechanic
169	.691	.812	Works well without a lot of explaining
123	.688	.799	On the lookout for better ways of doing the job
180	.688	.838	He gives his best effort to whatever he is doing
191	.688	.846	Sets a good example through his habits of work
163	.686	.834	Handles additional and special responsibilities well
92	.684	.834	He should develop into one of the leading men in the crew
145	.684	.808	Can take over in an emergency
179	.683	.807	Is determined to make good
200	.680	.784	Is always a willing worker
91	.671	.836	Very valuable in a new operation
106	.669	.791	Very good at going ahead alone
152	.667	.773	Respected by all who know him
144	.666	.745	Helps others with their work whenever possible
90	.664	.731	Selects proper Tech. Orders and makes correct use of them
111	.662	.797	He makes the best use of his ability
34	.657	.769	Uses good judgment in most situations
174	.657	.811	Knows how to organize to get right things done at right time
114	-.686	-.806	Loses too much time in waste motion
170	-.676	-.799	Never has things ready when needed
121	-.670	-.795	Always ready with an excuse
66	-.664	-.798	Seldom plans work ahead
65	-.659	-.824	Takes far too many breaks
88	-.659	-.797	Never notices things to be done
95	-.656	-.848	Doesn't do his share of the work

The criterion correlations are uniformly large for highly loaded items on this first square-root factor. The factor might perhaps be interpreted as the statistical representation of halo effects, and given the label of "General Success on the Job." Our reason for not doing so is that we believe this title is too broad, and does not make clear the psychological nature of this general factor. The term "General Job Efficiency" was chosen instead, as better representing the essential import of the factor. The listed items stress the mechanic's ability to organize his schedule so that his own work is coordinated with that of others and so that the more important jobs are done first. Such a mechanic can work by himself without detailed explanation or

constant supervision. His working habits are good, and we are told that he gives his best effort to whatever he is doing, a characteristic suggesting that his efficiency is a function of high motivation as well as of sound organization. The factor is perhaps more revealing in terms of what is omitted than what is included. It should be noted that little is said explicitly about the mechanic's manual dexterity or his knowledge of his job, and no reference is made to his social relations with the other mechanics, or to his emotional structuring. In this factor, the stress seems to be not on what the mechanic himself is like, but rather on his capacity for working efficiently and making good use of the skills at his command.

Factor 2. Social Maladjustment

Item number	Loading	Criterion correlation	Descriptive phrase
148	.769	-.736	Exaggerates too much
112	.386	-.703	Uses a "line" to try to impress others
146	.364	-.780	Wastes time walking around and talking to others
81	.356	-.653	Rubs people the wrong way
142	.351	-.673	Makes trouble over little things
184	.349	-.604	Always wanting sympathy
139	.324	-.483	He butts in on others' responsibilities
130	.321	-.752	Says one thing and does another
173	.319	-.507	Tries to flatter crew chief
200	-.318	.784	Is always a willing worker

This factor stresses personality inadequacies which result in unpopularity and generally poor interpersonal relations. The over-all flavor is one of immaturity,

self-centeredness, and obtrusiveness. It should be noted that the traits described tend to be correlated with disinclination for work.

Factor 3. Executive Ability

Item number	Loading	Criterion correlation	Descriptive phrase
166	.827	.621	Discusses plans frequently with superior
181	.316	.594	An extremely good organizer of personnel
163	.310	.834	Handles additional and special responsibilities well
40	.304	.635	Seeks and does additional tasks beyond those required
174	.286	.811	Knows how to organize to get right things done at right time
92	.284	.834	He should develop into one of the leading men of the crew
191	.280	.846	Sets a good example through his habits of work
165	.275	.612	Quick to praise a man for a job well done
122	-.282	-.568	Goes his own way regardless of others
136	-.275	-.621	Dissatisfied with his present job

Factor 3 defines the ability to organize activities and personnel and to spur co-operative achievement. The emphasis is not so much on the tendency to direct

and dominate others as on the capacity for responsible supervision and effective administration.

Factor 4. Leadership

Item number	Loading	Criterion correlation	Descriptive phrase
63	.822	.641	He is apt to be master of a situation if it needs a leader
21	.267	.505	Likes to take on responsibilities
189	.256	.523	He is good at organizing new activities
134	.255	.659	Plans ahead for almost any emergency
60	.253	.449	He makes definite decisions and defends them with courage
181	.251	.594	An extremely good organizer of personnel
4	.243	.364	Likes to make decisions
106	.212	.791	Very good at going ahead alone
182	-.271	-.407	Prefers to be a follower rather than a leader
67	-.208	-.795	Doesn't seem to realize he's just a cog in a large wheel

Factor 4 stresses decisiveness, initiative, and the tendency to assume a dominant role in leaderless situations. Here the emphasis is on liking for responsibility

and decision-making, that is to say, for giving directions rather than executing them. The factor appears to conform to traditional definitions of leadership.

Factor 5. Personal Charm

Item number	Loading	Criterion correlation	Descriptive phrase
107	.878	.363	Always pleasant and cheerful
71	.302	.171	Has a pleasant voice
76	.282	.352	Easy to talk to
176	.266	.460	He is considered a "good sport"
101	.251	.070	Easy going
9	.242	.340	Takes pride in his health
19	.239	.064	Usually sees the humorous side of things
80	.235	.472	Tends to see the good side of things
57	-.253	-.341	Is inclined to be impatient with others
93	-.235	-.376	Inclined to "pop off" on occasion

These items define a pleasant, relaxed, approachable personality with a pre-dominant tone of cheerfulness and good humor.

Factor 6. Resourcefulness

Item number	Loading	Criterion correlation	Descriptive phrase
158	.899	.434	Able to by-pass red tape when necessary
68	.230	.627	Is familiar with main trouble spots of an airplane
168	.226	.647	Will really go to bat for things when he has to
73	.222	.659	Usually finds some way to "crack a hard nut"
51	.216	.788	His advice is asked by others
45	.208	.649	Stimulates co-workers to think
177	.200	.372	Teaches himself how to operate equipment without supervision
33	.197	.721	He learns new jobs quickly
22	.183	.349	Has a good imagination
135	-.187	-.574	Is helpless when working in awkward positions

Ingenuity and resourcefulness appear to form the central core of this factor, with good job knowledge, superior learning ability, and an imaginative approach to problem-solving also associated.

Factor 7. Willingness and Adaptability

Item number	Loading	Criterion correlation	Descriptive phrase
140	.832	.504	Would be willing to begin in any kind of job to prove his ability
115	.201	.665	Wants to adjust himself to new situations
144	.186	.745	Helps others with their work whenever possible
179	.176	.807	Is determined to make good
107	.174	.363	Always pleasant and cheerful
28	.172	.705	Usually tackles a job with enthusiasm
119	.167	.611	Takes pride in group accomplishment
102	.164	.690	Turns off power or motor equipment when not in use
13	.160	.661	Can be used effectively on several different types of work
180	.153	.838	He gives his best effort to whatever he is doing

This factor emphasizes willingness, determination, and adaptability. The items portray the type of mechanic who wants to make good and prove his own capabilities. Eager to tackle new problems, he has high morale and takes pride in group accomplishment. It is hardly surprising to find that such an individual

often helps others. Though we cannot be sure as to the psychological reasons for this energetic service, one is inclined to speculate that the mechanics so described are by nature eager and compliant, and that it is this, rather than any overwhelming interest or ambition, which is responsible for their willingness.

Factor 8. Orderliness

Item number	Loading	Criterion correlation	Descriptive phrase
197	.806	.516	Knows his own abilities well
74	.217	.702	He accepts full responsibility for his own work
159	.168	.716	Usually assembles or disassembles parts in proper sequence
73	.160	.659	Usually finds some way to "crack a hard nut"
71	.158	.171	Has a pleasant voice
147	.157	.570	His quarters are usually neat
137	-.166	-.437	Is careless of personal appearance
10	-.165	-.564	Couldn't direct without irritating people
54	-.154	-.467	Hesitates in making decisions on ordinary everyday questions
135	-.150	-.574	Is helpless when working in awkward positions

At first sight, the items appearing under this factor are much more heterogeneous than those in the seven preceding it. We find here clustering together such varied characteristics as neatness, self-direction, insight, a pleasant voice, and accuracy in performance on the job. This diversity of item content can hardly be attributed to sampling error, as we find that factors with considerably less variance than this are nevertheless consistent and meaningful. We have accordingly hypothesized that this factor defines a personality trait making for precision and exactness in thought and

action. A mechanic having this trait knows his own mind and capacities, is neat and orderly in appearance, and takes a systematic approach to his work. He is methodical, but not without ingenuity, and can direct others without irritating them. In general, the impression given is of an individual who is well organized but not compulsive, careful but not finicky. This factor is an interesting one whose constituent items tend for the most part to correlate reasonably well with our criterion of job success.

Factor 9. Ability to Motivate Others

Item number	Loading	Criterion correlation	Descriptive phrase
171	.657	.715	Provides just the "spark" that is needed for effective teamwork
183	.209	.613	Makes allowance for the limitations of others
18	.204	.685	Arouses ambition in fellow mechanics
151	.192	.669	He has made a good social adjustment on the job
45	.190	.649	Stimulates co-workers to think
165	.187	.612	Quick to praise a man for a job well done
83	.185	.641	Boosts the morale of any group in which he is working
152	.179	.773	Respected by all who know him
189	.175	.523	He is good at organizing new activities
67	-.170	-.795	Doesn't seem to realize he's just a cog in a large wheel

Together these items describe the responsible, task-oriented mechanic with a talent for group activity. His attitude

toward the job and ability to organize make for effective teamwork among his co-workers.

Factor 10. Mechanical Proficiency

Item number	Loading	Criterion correlation	Descriptive phrase
73	.707	.659	Usually finds some way to "crack a hard nut"
34	.194	.760	Uses good judgment in most situations
159	.185	.716	Usually assembles or disassembles parts in proper sequence
61	.178	.728	He can organize his ideas effectively
68	.176	.627	Is familiar with main trouble spots of an airplane
13	.172	.661	Can be used effectively on several different types of work
33	.150	.721	He learns new jobs quickly
115	.158	.665	Wants to adjust himself to new situations
15	-.186	-.683	Doesn't listen to suggestions
95	-.172	-.848	Doesn't do his share of the work

In this factor the primary emphasis is upon proficiency and the ability to cope with mechanical problems. However, the psychological nature of this proficiency is defined only in the broadest of terms. Reference is made to ingenuity and good judgment, to versatility, adaptiveness, and, indirectly, to job knowledge. Motivational requirements are also mentioned, as they were in Factor 1. Supervisors do not seem to make the distinction between motivational and cognitive qualities which it is customary for psychologists to draw.

Orthogonality of the square-root factors. Since square-root factors are orthogonal, it may at first sight seem strange that several factors may be extracted in what might appear to be the same area. For example, Factor 3, labeled "Executive Ability" considers (among other things) ability to organize personnel, Factor 4, labeled "Leadership," stresses ability for giving directions and assuming a dominant role, and Factor 9, labeled "Ability to Motivate Others," describes a talent for group activity and for effective teamwork with co-workers. Moreover, some items appear in two lists, viz., "Quick to praise a man for a job well done," "An extremely good organizer of personnel," "He is good at organizing new activities," "He doesn't seem to realize he's just a cog in a large wheel."

Nevertheless, each factor has a somewhat different emphasis. Factor 3 stresses planning and organizational ability, Factor 4, leadership and the giving of orders, and Factor 9, cooperation and sympathy for others. Most of the items in each list are peculiar to that factor. These items which appear in two lists may be presumed to be factorially complex, embodying

some element of one factor and some element of another.

No doubt, by different placement of axes, some more comprehensive factor, extending over the entire area of personnel control and interpersonal relations, could be extracted, with subsidiary factors making the distinction between organizational ability, leadership, and cooperation. Any such solution would certainly be acceptable. Yet the way the variance has been partitioned by the square-root method seems equally acceptable in indicating the principal conceptual distinctions made by the supervisors in their descriptions.

An even more striking instance of apparent overlapping is between Factor 5, labeled "Personal Charm," and Factor 13, labeled "Changeability of Mood." Factor 13 not being reported in this monograph, it will suffice to note that the four items with highest loadings are: "Changeable in his attitudes," "Inclined to 'pop off' on occasion," "Gets mad if things don't go to suit him," and "He thinks he puts in too long hours." In contrast to the calm and relaxed personality of Factor 5, mechanics of the type portrayed by Factor 13 seem to be changeable, irritable, and given to outbursts of temper. But the cluster of negatively toned items in Factor 13 is not completely antithetical to the cluster of positively toned items in Factor 5. Factor 13 seems to be principally concerned with high emotionality, and Factor 5 with social manner. In the correlation matrix, the correlations between the two sets of items are well short of -1.00. This means that supervisors sometimes give the same response to both a positively toned item and a negatively toned item which appears to be antithetical. This may sometimes be the result of carelessness in responding to the inventory, but it seems reasonable to believe that no man's behavior is consistently good, poor, or average. This tendency to mark, say, "true" to antithetical phrases might on this line of reasoning be expected to be

especially characteristic of supervisors rating "average" mechanics. It may not be at all inconsistent for a supervisor to report a certain mechanic to be changeable in attitude, and yet to be pleasant to talk to and to possess a pleasant voice. These negative interitem phi coefficients of much less than unity probably provide evidence that the halo effect was to some extent overcome by the raters. The fact that items of the type included in Factors 5 and 13 tend to have lower criterion correlations supports this explanation. The fact that it is often difficult in our language to find exact synonyms and exact antonyms is a further indication of the complexity of our behavior.

V. RELATION OF THE FACTORS TO SUPERVISORY RATINGS OF JOB PROFICIENCY

In this section we consider: (a) the average criterion correlation for the ten most highly loaded items of each factor, (b) the pattern of job performance represented by items with high criterion correlations, and (c) the nature of the items which appear to be unrelated to ratings of either good or poor performance.

Average criterion correlations for the ten most highly loaded items for each factor. The criterion correlations will be recalled to be the phi coefficients between each item and the "best-poorest" classification, with average mechanics excluded from the sample. Since these judgments of "best" and "poorest" were made by the same supervisors who provided the item ratings, they do not represent an independent measure of job proficiency. The criterion correlations may therefore be spuriously high, both for this reason and because they are based on the extremes of the sample. The correlations are nevertheless useful for our present purpose, which is merely that of determining the pattern of behavior which supervisors consider to be characteristic of good mechanics.

Table 2 presents the average criterion correlations for the ten items with high-

TABLE 2
AVERAGE CRITERION CORRELATIONS
FOR THE FACTORS

Factor number	Factor title	Average criterion correlation
1	General job efficiency	.83
10	Mechanical proficiency	.71
3	Executive ability	.70
2	Social maladjustment	-.68
9	Ability to motivate others	.67
7	Willingness and adaptability	.66
6	Resourcefulness	.58
4	Leadership	.57
8	Orderliness	.54
5	Personal charm	.30

Note—Correlations are based on descriptions of "best" and "poorest" mechanics. ("Average" mechanics excluded.)

est loadings on each factor. In the two previous studies (6, 7), factors dealing with practical efficiency, good judgment, foresight, etc., had been found more predictive than those relating to sociability, character, degree of emotionality, and other personality attributes. Results for the Survey Inventory agree with those already reported. The most predictive factor is the first; this, it will be recalled, stresses practical efficiency and good organization. The least predictive factor is the one labeled "Personal Charm." It seems to make rather little difference to quality of work whether or not a mechanic is liked by his associates.

The Survey Inventory factors are, on the whole, more predictive than those for the other two inventories. This difference will be discussed in Section VI of this report.

The pattern of performance indicated by items with large criterion correlations. Tables 3 and 4 list items with criterion correlations greater than or equal to .70. Items more commonly used to describe "best" mechanics appear in Table 3; those more often applied to "poorest"

TABLE 3
CRITERION CORRELATIONS FOR ITEMS WITH HIGH POSITIVE PREDICTIVE VALUE ($\phi > .70$)

Item no.	Correlation coefficient	Item
187	.870	Fits well into the organization
116	.865	Goes over own work to see that it was done correctly
8	.857	Seems to have a good future as a mechanic
191	.846	Sets a good example through his habits of work
180	.838	He gives his best effort to whatever he is doing
91	.836	Very valuable in a new operation
92	.834	He should develop into one of the leading men in the crew
163	.834	Handles additional and special responsibilities well
161	.812	Works well without supervision
169	.812	Works well without a lot of explaining
174	.811	Knows how to organize to get right things done at right time
99	.809	Is a better man than the average we have had in past years
131	.809	Sees what will be needed later and gets ready for it
145	.808	Can take over in an emergency
179	.807	Is determined to make good
123	.799	On the lookout for better ways of doing the job
111	.797	He makes the best use of his ability
6	.795	Shows good foresight in his planning
106	.791	Very good at going ahead alone
51	.788	His advice is asked by others
200	.784	Is always a willing worker
86	.775	Does jobs in order of importance
152	.773	Respected by all who know him
34	.769	Uses good judgment in most situations
17	.749	Takes precautions to prevent damage to equipment
190	.748	Makes the best of every bad situation
144	.745	Helps others with their work whenever possible
90	.731	Selects proper Tech. Orders and makes correct use of them
41	.729	Has won nearly everybody's confidence
188	.729	Would be very difficult to replace
196	.729	Obtains co-operation from other people
27	.728	Checks on the condition of the airplane and its equipment prior to flight
61	.728	He can organize his ideas effectively
154	.723	Has exceptional mechanical ability
33	.721	He learns new jobs quickly
159	.716	Usually assembles or disassembles parts in proper sequence
171	.715	Provides just the "spark" that is needed for effective teamwork
70	.714	Makes plans fit in with those of fellow workers
36	.710	Readily understands and carries out instructions
28	.705	Usually tackles a job with enthusiasm
74	.702	He accepts full responsibility for his own work
24	.700	Knows how to put best foot forward

Note—Correlations are based on descriptions of "best" and "poorest" mechanics. ("Average" mechanics excluded.)

mechanics are presented in Table 4. The good mechanic will be seen to be described as:

(a) cooperative—he gets on well in the organization, is helpful to others, and succeeds in getting their cooperation;

(b) keen to do his best—he approaches a job enthusiastically, does the best with it that he can, and wants to make good;

(c) responsible—he works well without supervision, can take over in an emergency, is good

at going ahead by himself, and makes a success of any special responsibilities;

(d) versatile and flexible in his ways—he watches for better ways of doing the job, and is useful in new operations;

(e) accurate—he makes sure that his work is correct;

(f) foresightful with a good sense of proportion—he carries out jobs in the right order of importance, and gets ready for what is needed.

Some items in the list consider the ability level of the mechanic. The good

TABLE 4
CRITERION CORRELATIONS FOR ITEMS WITH HIGH NEGATIVE PREDICTIVE VALUE

Item no.	Correlation coefficient	Item
95	-.848	Doesn't do his share of the work
65	-.824	Takes far too many breaks
114	-.806	Loses too much time in waste motion
170	-.799	Never has things ready when needed
66	-.798	Seldom plans work ahead
88	-.797	Never notices things to be done
67	-.795	Doesn't seem to realize he's just a cog in a large wheel
121	-.795	Always ready with an excuse
146	-.780	Wastes time walking around and talking to others
53	-.761	He abuses privileges
79	-.759	His work requires more than a normal check-up
138	-.759	Does not take important matters seriously
192	-.759	Can't seem to make ends meet
47	-.756	Thinks mostly of himself and not of other crew members
130	-.752	Says one thing and does another
148	-.736	Exaggerates too much
1	-.729	Is frequently late for work
156	-.728	Is a routine worker, doing only what is required
109	-.718	Has a poor attitude
87	-.717	Unfairly criticizes policies of his superiors
186	-.716	Wastes fuels, lubricants and other supplies
118	-.709	Often picks up any old wrench to do a job
105	-.707	He criticizes men who are better than he is
167	-.706	Gripes about the least thing
112	-.703	Uses a "line" to try to impress others

Note—Correlations are based on descriptions of "best" and "poorest" mechanics. ("Average" mechanics excluded.)

mechanic is described as having high mechanical aptitude and being able to learn new jobs rapidly. There is specific reference to selecting appropriate Technical Orders and to assembling parts in proper sequence. Other items stress the practical application of knowledge and abilities.

It is interesting to notice that the good mechanic is rather generally respected for his work. We are told that he wins nearly everybody's confidence, and that others go to him for advice. Other items in the list describe successful job performance in such general terms that we are supplied with no new information as to its specific nature. Examples of such items (which are nevertheless highly predictive) are "He seems to have a good future as a mechanic" and "He would be very difficult to replace."

The items with high negative coefficients, i.e., those which are more characteristic of poor mechanics, reinforce the above picture. The poor mechanic is portrayed as lazy, careless, and uncooperative. He has low morale, fails to plan ahead, and is never ready on time. Untrustworthy and inefficient, he says one thing and does another, and loses time in waste motion. Perhaps most significantly of all, he is described as a routine worker who does only what is required and fails to take important matters seriously.

Items of low predictive value. Table 5 lists items with criterion correlations lower than .20. As in the earlier studies, the social personality of the mechanic bears little relation to the rating he is given by his supervisor. It matters little whether he is easygoing and at ease in

TABLE 5
CRITERION CORRELATIONS FOR ITEMS WITH LOW PREDICTIVE VALUE ($\phi < .20$)

Item no.	Correlation coefficient	Item
81	.064	Rubs people the wrong way
101	.070	Easy going
164	.121	Has a bold, direct manner
71	.171	Has a pleasant voice
16	.180	Is at ease in any situation
35	-.007	Often uses gestures to put across his ideas
100	-.046	Lets others take advantage of him
149	-.046	Is always asking for advice
185	-.088	Gets resentful if property or other rights are trespassed on
153	-.153	Worries over possible misfortunes
160	-.183	Fears criticism from superiors

Note—Correlations are based on "best" and "poorest" mechanics. ("Average" mechanics excluded.)

any social situation or is irritable and liable to worries and fears.

VI. COMPARISON OF THREE TYPES OF RATING ITEM

In the research series of which this study is a part, factor analyses were made of items drawn from three sources. The *Descriptive Inventory* (7) was derived from supervisors' interview descriptions of aircraft mechanics, the *Factorial Inventory* (6) from factor-analytic studies of personality and ability, and the *Survey Inventory*, reported here, from rating scales and questionnaires designed to measure job performance in various fields. In this section, results from the three studies will be compared.

The three inventories were administered to different samples of supervisors of aircraft mechanics under similar ex-

perimental conditions. The only material change was a decreased proportion of "average" ratings, relative to "best" and "poorest," in the Survey Inventory. The number and proportion of ratings of each type in each study are presented in Table 6. The increased proportion of extreme ratings in the Survey Inventory does not affect the criterion correlations, since they are based upon "best" and "poorest" ratings only. However, in the factor analyses there might be a greater concentration of item variance in the early factors of the Survey Inventory. This possibility should be kept in mind in the comparisons which follow.

Table 7 compares (a) the distribution of variance and (b) the average criterion correlation, for the first 10 factors of the three inventories. If, for the rating of job performance, we were obliged to re-

TABLE 6
COMPARISON OF SAMPLES IN THE THREE INVENTORIES

Inventory	Number of ratings				Percentage of ratings		
	Best	Average	Poorest	Total	Best	Average	Poorest
Descriptive	112	224	92	428	26.2	52.3	21.5
Factorial	95	194	94	383	24.8	50.7	24.5
Survey	148	179	137	464	31.9	38.6	29.5

TABLE 7

COMPARISON OF THE DISTRIBUTION OF VARIANCE
AND OF THE AVERAGE CRITERION CORRELATION IN
THE FACTORS OF THE THREE INVENTORIES

Inventory	Contribution to variance		Average criterion correlation	
	Factor 1	Factors 2-10	Factor 1	Factors 2-10
	per cent	per cent		
Descriptive	20.6	15.3	.74	.49
Factorial	9.0	23.4	.41	.41
Survey	26.8	17.0	.83	.60

Note—Correlations are based on "best" and "poorest" mechanics. ("Average" mechanics excluded.)

strict ourselves to items drawn from a single factor of a single inventory, the first factor of the Survey Inventory would have to be our choice. It defines better than any other the qualities considered by supervisors to be characteristic of a good mechanic. The criterion correlations are higher, and more item variance is accounted for, than for any other factor.

A rather different pattern emerges if the results for Factors 2-10 of each study are considered. The Factorial Inventory is seen to be more comprehensive in coverage than the Survey Inventory, but its items tend to be less job-centered and less predictive. The Descriptive Inventory falls between the other two in respect to predictive power.

The different areas of behavior described by each inventory. Differences in the subject matter of the inventories may perhaps be summarized most readily by making an a priori classification of the factors, and then attempting to determine the extent to which each behavioral area is represented in each inventory. The statistical orthogonality of the square-root factors need not preclude us

from doing this, since the presence of factorially complex items in several factors indicates their semantic interrelatedness. (See the discussion of orthogonality and overlapping of factors in the later part of Section IV.)

Because of the multiplicity of factors, the listing will be restricted to the ten largest factors in each study. Table 8 presents this classification. The grouping of factors is to some extent arbitrary, and alternative classifications are possible. Results are so clear, however, as to leave little doubt as to the nature of the differences in content.

1. *Practical ability.* By "practical ability," we mean the capacity to make decisions quickly, carry out jobs in order of importance, find better ways of doing a job, explain matters adequately, organize well, and plan work in advance. This aspect of job proficiency takes a central position in each inventory. Seven factors in this area appear in Table 8, and criterion correlations are in most cases high.

2. *Sense of responsibility toward work.* There are four factors in this area in the Descriptive Inventory, and one in the Survey Inventory. A mechanic's willingness and dependability seem to be the aspect of job behavior which is discussed at greatest length by the supervisors. As might be predicted, criterion correlations in this area are high.

3. *Pattern of interests.* Interest in the job is not well represented in any of the three inventories. The Descriptive Inventory contains a factor entitled "Interest in Aircraft Maintenance," and the Survey Inventory factor of "Willingness and Adaptability" might perhaps also have been included in this grouping. In both factors, the items are rather broad and nonspecific. The Factorial Inventory

TABLE 8
COMPARISON OF THE TEN MAJOR FACTORS IN THE DESCRIPTIVE, FACTORIAL,
AND SURVEY INVENTORIES

Area of behavior	Factor number	Average criterion correlation	Factor title
1. Practical ability	D6	-.71	Failure to use knowledge effectively
	D7	.41	Teaching capacity
	F3	-.71	Tendency to indecision
	F4	.58	Practical ability
	S1	.83	General job efficiency
	S3	.68	Executive ability
	S10	.71	Mechanical proficiency
2. Sense of responsibility towards work	D1	.74	Sense of responsibility
	D3	.50	Willingness for work
	D4	-.59	Laziness and lack of initiative
	D8	.58	Industriousness
	S7	.66	Willingness and adaptability
	D2	.41	Interest in aircraft maintenance
3. Pattern of interests	F6	.48	Cultural interests
4. Intellectual ability	D9	.55	Memory
5. Personality characteristics in the job situation	F5	-.58	Lack of cooperation
	S6	.58	Resourcefulness
	S8	.54	Orderliness
	S4	.57	Leadership
	S9	.67	Ability to motivate others
	F2	-.62	Social immaturity
6. Social adjustment	F7	-.06	Shyness
	F8	.16	Sociability
	F9	.18	Submissiveness
	S2	-.68	Social maladjustment
	S5	.30	Personal charm
	F1	-.41	Tendency to neuroticism
7. Emotional adjustment	F10	-.31	Emotional sensitivity
8. Character	D5	-.38	Weakness of character
	D10	.23	Self-control

Notes—1. Correlations are based on "best" and "poorest" mechanics. ("Average" mechanics excluded.)

2. D, F, and S indicate that the factors are from the Descriptive, Factorial, and Survey Inventories respectively.

contributes a factor of "Cultural Interests" to this area. This serves to indicate the different and much less job-relevant type of item utilized by factor-analysts in their studies of personality and ability.

4. *Mental abilities.* Only slight reference is made to ability and to job knowledge. In the report on the Descriptive Inventory (7, p. 232), we commented on the fact that supervisors stressed motivation and practical efficiency, so that no clearly defined factors of memory, reasoning, and problem-solving ability

emerged. Items were included in the Factorial Inventory to assess mental abilities. Factors of creativity, mental alertness, etc., emerged, and the results leave no doubt that all-round intellectual dullness is a variable contributing to poor performance on the job. However, these factors all account for a rather small proportion of the variance. This means that the differences between mechanics are mostly described in such a way as to be attributable to factors of a more practical kind. Items dealing with sharpness of mind, ability to learn quickly,

etc., tend to appear along with items describing general job efficiency, organizational capacity, and the like, rather than by themselves. In no study is there extended reference to a mechanic's knowledge of his job.

5. *Personality characteristics in the job situation.* Personality traits seem to divide into two groups: (a) those which are relevant to the job situation, such as being orderly, resourceful, and able to motivate others; (b) those which are only slightly or not at all relevant in the job situation, such as degree of shyness. There are five factors under the heading of "Personality characteristics in the job situation." Four of these are from the Survey Inventory. Probably because the items for this inventory were written with specific reference to the job situation, the stress in the personality-type items is upon degree of cooperativeness, ability as a leader, etc.

6. *Social adjustment.* The Factorial Inventory items deal to a greater extent than the others with the entire pattern of a man's interpersonal relationships, and his ability to manage adequately in social situations. Four of the six factors in this area appear in the Factorial Inventory. Criterion correlations are in general lower for these factors. The shy and introverted mechanic seems to be quite acceptable. On the other hand, the hostile and trouble-making mechanic is not.

7. *Emotional adjustment.* The two factors in this area are, as might be expected, from the Factorial Inventory.

8. *Character.* The Descriptive Inventory makes greater reference than any other to defects of character—heavy drinking, irresponsibility in money matters, and the like. Criterion correlations are, as a rule, rather low.

From Table 8, a summary statement can be made of the distinctive characteristics of each of the three inventories.

The Descriptive Inventory (phrases supplied by supervisors) stresses a mechanic's willingness for work, dependability, and sense of responsibility. Furthermore, there is more in this inventory than in either of the others about a mechanic's moral habits—i.e., about his monetary and drinking habits, his trustworthiness, his punctuality, etc. These are probably areas of behavior where there are likely to be disputes between supervisor and mechanic, so that any deficiencies are strongly impressed upon the supervisor's attention. The supervisors tend to use evaluative rather than analytical descriptions.

The Factorial Inventory (phrases drawn from factor-analytic studies reported in the literature) stresses the mechanic's social and emotional adjustment—i.e., his emotional sensitivity and variability, his shyness, tendencies to submissiveness, etc. In general, these aspects of personality are not very closely related to job performance. The criterion correlations are lower than for any other area of behavior. This finding must be qualified in one important respect. There are factors of "social immaturity" and "social maladjustment" which portray "difficult" and somewhat antisocial personalities (as opposed to merely shy and submissive individuals), and these mechanics are rated poor. In addition, those factors which relate more directly to social behavior on the job itself (e.g., leadership ability, ability to obtain cooperation from others) have rather high criterion correlations.

The Survey Inventory (phrases obtained from researchers in applied psy-

8-366
7.9.70

chology) stresses practical ability. It also yields factors which describe what we have called "Personality characteristics of the job situation," viz., resourcefulness and orderliness. Its factors of social adjustment are rather strongly job oriented. Motivation is stressed less in the Survey Inventory than in the Descriptive Inventory, but there is greater emphasis upon efficiency and good interpersonal relations. The Survey Inventory items are less evaluative but more analytic than those of the Descriptive Inventory, and more job centered than those of either of the other inventories.

This greater "job-centeredness" of the Survey Inventory is reflected in its appreciably higher criterion correlations. Thirty-three per cent of the Survey Inventory criterion correlations are greater than .70. Only 18 per cent of the Descriptive Inventory items and 10 per cent of the Factorial Inventory items are as high as this. Hence, if we had to confine ourselves to any single one of the three inventories, the Survey Inventory would be the most reasonable choice.

The over-all picture of the good mechanic presented by the three studies is that of a man who has in mind the needs and demands of his job, who is willing to work hard, and who is vigorous and mentally alert, displaying good judgment and powers of decision. He is precise, orderly, and dependable; yet at the same time he displays initiative and acts for himself. He is efficient and practical. *More is said about the use he makes of his abilities than about the abilities themselves. The qualities which have been listed are probably useful assets in any form of work.* Little is said about the special features of a mechanic's job situation which distinguish it from other types of work.

One of the important determinants of quality of performance on the job seems to lie in differing standards of value: we can differentiate between the mechanic who cares about his work, is ambitious, and wants to get on, and the one whose interest lies in getting privileges and time off and avoiding work. A second main line of distinction is between the mechanic who is systematic and orderly and his associate who is inefficient and lacking in common sense. In general, these items are strongly job-centered, and the occupational characteristics which they describe are not always easily expressed in the standard terms used for accounts of ability and personality structuring.

The dependence of factorial results upon the sampling of variables. The results of the three studies show very clearly how content differs from one inventory to another. Obviously selection of items in factorial studies of this type is a matter of the utmost experimental importance. Before any final and definitive factorial description of the area can be provided, *some logical definition is needed of a population of items from which random samples of items may be drawn.* Once the population of variables can be determined, then the correlational configuration will tend to be similar from one factorial study to another (provided the sample of variables was large, and the persons tested in the various studies also formed random samples from a defined population). Given this similarity of the correlational configuration, the clusters obtained from one random sample of items will then tend to agree with the clusters obtained from another random sample.

The problem then is in the definition of the population of variables. There

seem to be at least two requirements: (a) to establish the possibility of enumeration of all members of the population or of an unbiased and sufficiently large proportion of the members of the population to provide a basis for random sampling, (b) to demonstrate that the population so defined provides adequate representation of the area of behavior to be investigated.

Whenever we are operating with environmental givens (i.e., variables already present in the psychological environment, as opposed to artifacts constructed at will by the investigator), it is generally not difficult to formulate some criterion for inclusion in the population, and accordingly to devise some random basis for sampling. For example, current semantic usage can provide a basis for sampling. The complete set of adjectives appearing in an English dictionary can be listed, as has been done by Allport and Odbert (1), and the sample drawn from these. Cattell (3, p. 216) prepared a comprehensive list of "basic trait elements" in this way. Likewise environmental situations can be sampled. Brunswik and his co-workers (2, pp. 41-46), after protesting that many laboratory experiments use unimportant variables, unrepresentative of the environment and interlocked in some artificial way, have made effective use of sampling various environmental situations in their studies of size constancy.

Our procedure in selecting items for the three inventories was somewhat similar. For the Descriptive Inventory (7) phrases were sampled from the descriptions of mechanics made by their supervisors. The population in this case consists of the various phrases currently used by Air Force supervisors to describe the mechanics under their control. The

Factorial Inventory (6) was developed from the population of items used in published factor analyses in the area of personality, and the Survey Inventory from the population of published items in rating scales and questionnaires in industrial psychology.⁹

The principal difficulty is in meeting the second requirement for a population of variables. No one of our three proposed definitions of the population of job-knowledge descriptions proved to be sufficiently broad to subsume all aspects of job performance. For example, reasons for interest or disinterest in work are not adequately elicited by any inventory. Likewise, no one provides a sufficiently explicit account of the role played by job knowledge in job performance. To study these areas, further items would have to be prepared. *Evidently the task of item-writing is not yet at an end, despite the large number of scales prepared by psychologists during the last fifty years.*

This appearance of precise definition of the population may be illusory even when operating with the population of dictionary adjectives, as Cattell did. Many adjectives have various shades of meaning, and often no instructions are given to the subjects requiring them to select one of these meanings rather than any other. Indeed, meaning can generally be better indicated by phrases, which provide a context, than by words in isolation.

Once, however, we are obliged to conclude that there is a place for further

⁹ The sampling in this latter case was modified by the attempt to tap previously unrepresented areas of behavior, but the general argument is not thereby affected, since further inventories similar to the Survey Inventory could obviously be obtained from the original collection of items.

item-writing by the investigator, then it follows that we are at present only partially able to achieve a definition of a population of items. So long as the investigator is free to create new items at will, it seems to be premature to try to provide a rigorous delimitation of the population of items from which a sample is to be drawn.

The function of the factor analysis. Because we are not yet in a position to define sampling procedure exactly, the development of a satisfactory rating scale for job performance has therefore to depend upon the progressive selection, refinement, and supplementation of items. The principal function of the factor analysis reported here has been to provide an empirical basis for preparing a briefer scale. The interview situation can then be used to assist in elucidation of the factors represented therein. After describing a mechanic in terms of the items, a supervisor can be asked further questions, e.g., with respect to the reasons for unusually high motivation, whether it indicates a man's interest in his own advancement, or his liking for mechanical assignments, or a favorable attitude toward the Air Force and its organization. These supplemental questions can be framed much more appropriately around the structure of the factors than with respect to individual items. The factor analyses, therefore, represent an initial step toward the better definition of the population of items for the rating of mechanics' performance.

VII. SUMMARY

The research here reported is the last of three studies designed to investigate the different psychological correlates of job proficiency in the field of aircraft mechanics. Inventory items hypothesized to be related to job performance in this

field were assembled from various sources, and supervisors were required to describe mechanics in terms of these phrases. This report considers the factorial structure of the items, and the relation of each item to the supervisors' over-all ratings of the mechanics.

The report is also designed to provide an example of application of the square-root method of factor analysis, as modified by Wrigley and McQuitty (9). The method has been devised for use with correlation matrices which are too large to be analyzed by standard procedures. There is a brief outline of the criticisms generally made of the square-root method, and the modifications introduced in an attempt to overcome these deficiencies.

Items in present study were systematically drawn from a set of rating scales and questionnaires which had been designed by various psychologists to measure different aspects of job proficiency. The items in the initial sample were revised in accordance with the results of a preliminary administration. The revised scale of 200 items, known as the Survey Inventory, was then administered to 464 supervisors of Air Force Aircraft and Engine Mechanics. They were required to describe a "best," "poorest," or "average" mechanic of their own choosing in terms of "true-false" responses to the items.

A square-root factor analysis was then made of the correlations between these item responses. The ten leading factors were named: (a) General Job Efficiency, (b) Social Maladjustment, (c) Executive Ability, (d) Leadership, (e) Personal Charm, (f) Resourcefulness, (g) Willingness and Adaptability, (h) Orderliness, (i) Ability to Motivate Others, (j) Mechanical Proficiency.

Phi coefficients were also calculated be-

tween the inventory items and the "best-poorest" dichotomy. These correlations indicate the factors and the individual items which supervisors judge to be most relevant to job proficiency, and thereby serve to define more precisely supervisors' conceptions of the nature of good and poor job performance.

These results were then compared with those obtained from the analysis of (a) the Descriptive Inventory (7) (where items were derived from relatively unstructured descriptions of "best" and "poorest" mechanics by their supervisors) and (b) the Factorial Inventory (6) (where items were derived from factor-analytic studies). The inventories agree in the findings that practical capabilities are more necessary in a successful mechanic than general intellectual ability, and that lack of motivation and a poor sense of responsibility are more detri-

mental than a poor level of socioemotional adjustment. Each inventory tends, however, to have a distinctive coverage of its own. The Descriptive Inventory stresses a mechanic's willingness for work and his dependability, the Factorial Inventory his social and emotional adjustment, and the Survey Inventory, his practical ability. The items of the Survey Inventory are shown to be of broader coverage and somewhat more predictive and job-centered than those in the two preceding inventories. Hence the Survey Inventory would be the most reasonable choice if a single one of the three were to be selected for assessing job proficiency. Because of the distinctive nature of each inventory, however, a shortened scale may be expected to be a better measuring instrument if based on all three inventories than if developed from any single one.

APPENDIX

SURVEY INVENTORY FOR AIR FORCE MAINTENANCE PERSONNEL

PURPOSE

This inventory is part of a research study on the attitudes and work habits of Airplane and Engine Mechanics. Your assistance is needed in carrying out this research because of your experience and close association with men in the maintenance field.

INSTRUCTIONS

Think back over the crews of Airplane and Engine Mechanics you have worked with during the last two years. Think of a person who is, in your opinion, just an *AVERAGE*¹⁰ mechanic; that is, a mechanic who is neither especially good, nor especially poor. Do *not* pick out a crew chief or supervisor when you make this selection.

Statements are listed on the following pages which will help you describe the mechanic you

now have in mind. Say each statement to yourself and decide whether or not it fits this mechanic. If it does, circle the T (for true) at the left of the statement; if not, circle the F (for false).

Decide on each statement before going on to the next. *Guess* whenever you are not certain whether or not a statement describes the man. Be sure to answer all statements as either true or false.

NOTE: All information will be held in strictest confidence and used for research purposes only. It WILL NOT affect you or the mechanic in any way.

SURVEY INVENTORY

- | | | |
|---|---|--|
| T | F | 1. Is frequently late for work |
| T | F | 2. Is unwilling to lend a helping hand |
| T | F | 3. Has a very sharp mind |
| T | F | 4. Likes to make decisions |
| T | F | 5. Talks a lot to everybody |
| T | F | 6. Shows good foresight in his planning |
| T | F | 7. Lacks snap in getting the work done |
| T | F | 8. Seems to have a good future as a mechanic |

¹⁰ Other supervisors received directions to describe the "Best" mechanic they had known during the previous two years, and still others, to describe the "Poorest" mechanic.

- | | | | |
|-----|---|-----|---|
| T F | 9. Takes pride in his health | T F | 50. Makes promises he doesn't keep |
| T F | 10. Couldn't direct without irritating people | T F | 51. His advice is asked by others |
| T F | 11. He remains calm under pressure | T F | 52. Apparently not physically fit |
| T F | 12. Has slow reactions | T F | 53. He abuses privileges |
| T F | 13. Can be used effectively on several different types of work | T F | 54. Hesitates in making decisions on ordinary everyday questions |
| T F | 14. Lacks confidence in own opinions | T F | 55. Frequently gives short answers |
| T F | 15. Doesn't listen to suggestions | T F | 56. Has trouble adjusting himself to changed conditions |
| T F | 16. Is at ease in any situation | T F | 57. Is inclined to be impatient with others |
| T F | 17. Takes precautions to prevent damage to equipment | T F | 58. Uses parts catalogue and stock list correctly in ordering parts |
| T F | 18. Arouses ambition in fellow mechanics | T F | 59. Lets difficulties get him down |
| T F | 19. Usually sees the humorous side of things | T F | 60. He makes definite decisions and defends them with courage |
| T F | 20. He thinks he puts in too long hours | T F | 61. He can organize his ideas effectively |
| T F | 21. Likes to take on responsibilities | T F | 62. Expects too rapid advancement |
| T F | 22. Has a good imagination | T F | 63. He is apt to be master of a situation if it needs a leader |
| T F | 23. Criticizes co-workers | T F | 64. Needs to be prodded occasionally |
| T F | 24. Knows how to put best foot forward | T F | 65. Takes far too many breaks |
| T F | 25. Is excellent for instructing beginners | T F | 66. Seldom plans work ahead |
| T F | 26. His personal conduct cannot be criticized | T F | 67. Doesn't seem to realize he's just a cog in a large wheel |
| T F | 27. Checks on the condition of the airplane and its equipment prior to flight | T F | 68. Is familiar with main trouble spots of an airplane |
| T F | 28. Usually tackles a job with enthusiasm | T F | 69. He thinks his supervisors are too strict |
| T F | 29. Conforms to regulations and policies | T F | 70. Makes plans fit in with those of other fellow workers |
| T F | 30. Does not improve with experience | T F | 71. Has a pleasant voice |
| T F | 31. Sometimes acts childish | T F | 72. Forgets to return borrowed tools |
| T F | 32. Never quits ahead of time | T F | 73. Usually finds some way to "crack a hard nut" |
| T F | 33. He learns new jobs quickly | T F | 74. He accepts full responsibility for his own work |
| T F | 34. Uses good judgment in most situations | T F | 75. Allows his emotions to color his judgment |
| T F | 35. Often uses gestures to put across his ideas | T F | 76. Easy to talk to |
| T F | 36. Readily understands and carries out instructions | T F | 77. Gets mad if things don't go to suit him |
| T F | 37. Has about reached limit of his abilities | T F | 78. Quick to size up a situation |
| T F | 38. Will trample over anyone to benefit himself | T F | 79. His work requires more than a normal check-up |
| T F | 39. Usually quick and accurate in emergencies | T F | 80. Tends to see the good side of things |
| T F | 40. Has poor posture | T F | 81. Rubs people the wrong way |
| T F | 41. Has won nearly everybody's confidence | T F | 82. He seldom asks for any time off |
| T F | 42. Often takes rumors seriously | T F | 83. Boosts the morale of any group in which he is working |
| T F | 43. Plenty of military snap and bearing | T F | 84. Is very selfish |
| T F | 44. Doesn't clean tools before putting them away | T F | 85. Uses correct torque on bolts and nuts |
| T F | 45. Stimulates co-workers to think | T F | 86. Does jobs in order of importance |
| T F | 46. Can stand kidding | T F | 87. Unfairly criticizes policies of his superiors |
| T F | 47. Thinks mostly of himself and not of other crew members | T F | 88. Never notices things to be done |
| T F | 48. Does very little original thinking | T F | 89. Feels that he owes nothing to anybody |
| T F | 49. Seeks and does additional tasks beyond those required | T F | 90. Selects proper Tech. Orders and makes correct use of them |

- T F 91. Very valuable in a new operation
T F 92. He should develop into one of the leading men of the crew
T F 93. Inclined to "pop off" on occasion
T F 94. Risks getting in bad with superiors to stand up for his fellow workers
T F 95. Doesn't do his share of the work
T F 96. Can't be told anything
T F 97. Can talk intelligently on almost any topic
T F 98. Cannot handle several details of his job at the same time
T F 99. Is a better man than the average we have had in past years
T F 100. Lets others take advantage of him
T F 101. Easy going
T F 102. Turns off power or motor equipment when not in use
T F 103. Is usually dressed neatly
T F 104. Gets help when in difficulties
T F 105. He criticizes men who are better than he is
T F 106. Very good at going ahead alone
T F 107. Always pleasant and cheerful
T F 108. He always waits his turn
T F 109. Has a poor attitude
T F 110. Always shows proper respect toward superiors
T F 111. He makes the best use of his ability
T F 112. Uses a "line" to try to impress others
T F 113. Resents being given a rush job
T F 114. Loses too much time in waste motion
T F 115. Wants to adjust himself to new situations
T F 116. Goes over own work to see that it was done correctly
T F 117. His appearance creates a distinctly favorable impression
T F 118. Often picks up any old wrench to do a job
T F 119. Takes pride in group accomplishment
T F 120. Has "jumpy" nerves
T F 121. Always ready with an excuse
T F 122. Goes his own way regardless of others
T F 123. On the lookout for better ways of doing the job
T F 124. Changeable in his attitudes
T F 125. Not as tactful in criticizing as he should be
T F 126. Is not very sociable
T F 127. Does everything possible to keep maintenance costs down
T F 128. Avoids doing things which do not reflect credit on him
T F 129. Makes mistakes in judging the character and ability of others
T F 130. Says one thing and does another
T F 131. Sees what will be needed later and gets ready for it
T F 132. Is a never-tiring worker
T F 133. Never cocky or over-bearing
T F 134. Plans ahead for almost any emergency
T F 135. Is helpless when working in awkward positions
T F 136. Dissatisfied with his present job
T F 137. Is careless of personal appearance
T F 138. Does not take important matters seriously
T F 139. He butts in on others' responsibilities
T F 140. Would be willing to begin in any kind of job to prove his ability
T F 141. Resents criticisms or suggestions
T F 142. Makes trouble over little things
T F 143. Reports on time for scheduled appointments, meetings, formations, etc.
T F 144. Helps others with their work whenever possible
T F 145. Can take over in an emergency
T F 146. Wastes time walking around and talking to others
T F 147. His quarters are usually neat
T F 148. Exaggerates too much
T F 149. Is always asking for advice
T F 150. Always does things in one particular way
T F 151. He has made a good social adjustment on the job
T F 152. Respected by all who know him
T F 153. Worries over possible misfortunes
T F 154. Has exceptional mechanical ability
T F 155. Often waits unnecessarily for directions
T F 156. Is a routine worker, doing only what is required
T F 157. He is quarrelsome
T F 158. Able to by-pass red tape when necessary
T F 159. Usually assembles or disassembles parts in proper sequence
T F 160. Fears criticism from superiors
T F 161. Works well without supervision
T F 162. He seems to know when to keep his mouth shut
T F 163. Handles additional and special responsibilities well
T F 164. Has a bold, direct manner
T F 165. Quick to praise a man for a job well done
T F 166. Discusses plans frequently with superior
T F 167. Gripes about the least thing
T F 168. Will really go to bat for things when he has to

- T F 169. Works well without a lot of explaining
- T F 170. Never has things ready when needed
- T F 171. Provides just the "spark" that is needed for effective teamwork
- T F 172. A one-job man, no wide grasp of maintenance
- T F 173. Tries to flatter crew chief
- T F 174. Knows how to organize to get right things done at right time
- T F 175. Thinks his work is too monotonous
- T F 176. He is considered a "good sport"
- T F 177. Teaches himself how to operate equipment without supervision
- T F 178. Complaints are made (more often than average) on the quality of his work
- T F 179. Is determined to make good
- T F 180. He gives his best effort to whatever he is doing
- T F 181. An extremely good organizer of personnel
- T F 182. Prefers to be a follower rather than a leader
- T F 183. Makes allowance for the limitations of others
- T F 184. Always wanting sympathy
- T F 185. Gets resentful if property or other rights are trespassed on
- T F 186. Wastes fuel, lubricants, and other supplies
- T F 187. Fits well into the organization
- T F 188. Would be very difficult to replace
- T F 189. He is good at organizing new activities
- T F 190. Makes the best of every bad situation
- T F 191. Sets a good example through his habits of work
- T F 192. Can't seem to make ends meet
- T F 193. Unwilling to stick his neck out
- T F 194. Inclined to pay much attention to his aches and pains
- T F 195. He likes the way the maintenance squadron is run
- T F 196. Obtains co-operation from other people
- T F 197. Knows his own abilities well
- T F 198. He accepts decision of the majority without showing his displeasure when the decision is against him
- T F 199. Fails to make a good first impression
- T F 200. Is always a willing worker

REFERENCES

1. ALLPORT, G. W., & ODBERT, H. S. Trait-names. A psycho-lexical study. *Psychol. Monogr.*, 1936, 47, No. 1. (Whole No. 211).
2. BRUNSWIK, E. *Systematic and representative design of psychological experiments*. Berkeley: Univer. of California Press, 1949.
3. CATTELL, R. B. *The description and measurement of personality*. Yonkers-on-Hudson, N.Y.: World Book Co., 1946.
4. GUTTMAN, L. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 1940, 5, 75-99.
5. HOLZINGER, K. J., & HARMAN, H. H. *Factor analysis*. Chicago: Univer. of Chicago Press, 1941.
6. MCQUITTY, L. L., CROSS, K. PATRICIA, & WRIGLEY, C. F. Approaches to the prediction of job success: II. The Factorial Inventory. Unpublished manuscript on file in the Univer. of California Library, 1954.
7. MCQUITTY, L. L., WRIGLEY, C. F., & GAIER, E. L. An approach to isolating dimensions of job success. *J. appl. Psychol.*, 1954, 38, 227-232.
8. STAUGAS, L., & MCQUITTY, L. L. A new application of forced-choice ratings. *Personnel Psychol.*, 1950, 3, 413-424.
9. WRIGLEY, C. F., & MCQUITTY, L. L. The square root method of factor analysis: a re-examination and a shortened procedure. *USAF Person. Train. Res. Bull.*, in press.

(Accepted for publication February 25, 1956)

Psychological Monographs: General and Applied

The Effect of Varying Amounts and Kinds of Information
As Guidance in Problem SolvingBERNARD R. CORMAN¹*Michigan State University*

I. INTRODUCTION

A PROBLEM may be said to exist when habitual responses fail to lead to the attainment of a desired goal. A search for a new response must then be undertaken. In intellectual problems, the problem solver must not only search for possible alternative courses of action, but must also select from among these alternatives the ones that will most successfully remove the obstacle to the goal. The search, then, is for information that will give structure to the problem. Presumably, as the amount of information developed is increased, the necessity for search is reduced. And, when all relevant information for the problem is available and understood, a *problem* no longer exists, although practice may still be required to make the new response habitual.

Information may be developed by the problem solver's own motivated search, or it may be made available to him. In-

formation supplied by directions, by clues, or by other guidance, however, may vary in amount. There may be a greater or lesser reduction in the number of alternatives which remain to be discriminated.

If little or no information is supplied, the problem solver may fail to develop the primary information needed. This failure, in turn, may lead to frustration and a turning away from the problem. It has been shown that, left to their own devices, problem solvers often set up false assumptions which make for unnecessary restrictions that delay or prevent solution (14). Those who argue the efficacy of specifically directive guidance, point to the large amount of time expended in the discovery of the required information, with the important risk of the development and perseveration of errors (21, p. 147).

On the other hand, by presenting large and significant amounts of information, the necessity for search on the part of the problem solver may be lessened. The opportunity for a thorough examination of the problem may thereby be restricted. Situations have been described in which specific directive information was accompanied by failure to solve problems like those on which the guidance had been given (9, 12, 23). This lack of transfer has been explained as a failure of the solver to become cognizant of

¹ This study is a revision of a dissertation submitted in 1954 to the faculty of Teachers College, Columbia University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The writer is deeply indebted to Dr. Irving Lorge, who directed the dissertation, for his sustained interest, encouragement, and stimulating counsel. The writer also wishes to express his appreciation to Drs. Robert Thorndike, Herbert Solomon, and Lincoln Moses for their many helpful suggestions.

relationships among the structural elements of the problems studied. Information supplied in guidance, it has been suggested, should provide an opportunity for discovery of these relationships by the problem solver through his own motivated efforts, rather than anticipate such discovery by the identification of specific methods or principles (9). While in "discovery" more errors may occur, it has been felt that, in the context of a more intense search, such errors would not be deleterious (19).

Faced with two seemingly contradictory points of view, those charged with the guidance of learning experiences in problem solving are asked to choose between two unattractive alternatives in determining how much information to provide. It is the purpose of this monograph to re-examine the effects of giving varying amounts of information in the guidance of problem solving.

II. RATIONALE FOR AN EXPERIMENT

If problem solving (as distinguished from routine drill) involves the search for a new response, then some "discovery" will accompany the successful solution of any problem. As Stroud has observed:

The case for discovery is difficult to evaluate, and indeed cannot be evaluated until the term is given precise definition. There is discovery with help, much or little, and discovery without help. In a sense any act of understanding is a discovery no matter by how much help or explanation the understanding is achieved . . . (20, pp. 582-583).

It follows that, while the extent of search may be more or less restricted, any test of the effectiveness of varying amounts of information given as guidance would require that, as an experimental condition, *all* subjects (*Ss*) be engaged in discovering a correct response.

In many of the experiments which have been taken as supporting the desirability of methods of tuition emphasizing "discovery" the comparisons have been between groups who have been shown a correct solution and groups who must discover such a solution (9, 13, 19). Such studies are more correctly interpreted as giving additional support for methods of tuition emphasizing "understanding" rather than methods of routine drill. The question to be answered, however, is whether "understanding" itself, as measured by success in transfer, is enhanced by greater or lesser amounts of information given as guidance.

A distinction must also be made between "amount of information" and "number of clues." It would be possible to design an experiment in which contrasted groups receive varying numbers of clues. The clues, moreover, might differ in the amount of information each made available (that is, in the extent to which the *S* was enabled to discriminate between possible alternative courses of action) (4). Whatever the discrimination permitted by the least explicit of the clues, this discrimination would also be made with each additional and more direct clue. The effect of such repetition would be to reduce errors and increase understanding quite apart from the fact that with each additional clue, additional information might be provided. It would follow that, where the effectiveness of differing amounts of information is the focus of interest, a necessary experimental control would be that the information any particular *S* receives be the same information at each repetition of clues, and that the number of repetitions be the same for all *Ss*.

A number of experiments have been reported where the two experimental

safeguards described above were given consideration. Typically, in these studies, Ss given a general principle, underlying a related cluster of problems, were less successful than were Ss whose attention was simply called to certain structural features of the problem situations (5, 12, 16, 22, 23). Since the application of the generalization should lead directly to solutions, these results have also been taken as supporting evidence for the superiority of methods of tuition allowing for greater search and "discovery." But an alternative explanation is possible which would take account of the observation that information may differ in kind as well as in amount.

Duncker has suggested that problems are usually attacked in stages, the solver setting up a series of subproblems (5). Protocols of problem-solving experiments would indicate that, while a few individuals may begin by trying to discover a generalization which may, when found, be applied to all variations of the problem, the more typical process is to attempt first to discover a method of attacking and solving specific examples (6, 18). Generalization may or may not follow success with a number of problem variations. If this is so, then the greater effectiveness of information related to a method of attacking specific examples of a problem-type may be explained as arising not from any inherent difference in the awareness or activity demanded of the S, but from the appropriateness of the information to the task the problem solver has set himself. Information given about the principle involved may be quite inappropriate in the early stages of problem solving.

If guidance toward the discovery of an attack on a problem (method informa-

tion) and information about a general principle underlying a cluster of similar problems (rule information) are, in fact, two *kinds* of information relating to two different aspects of problem solving, then not one, but two, criteria of success of guidance would be indicated; viz., success in solving the problems, and success in verbalizing the rule. Where the information made available is appropriate to the criterion used, then the guidance should prove beneficial. Where the information is inappropriate, guidance should prove ineffective, if not actually detrimental.

Further, if information about rule and about method are two kinds of information, it is possible to envisage guidance that would make available to the student greater or lesser *amounts* of either kind of information. If the measure of effectiveness is appropriate to the kind of information made available, and if the less explicit information proves more effective, then additional support would be given to the position that opportunity for greater self-directed search by the student is desirable as guidance. If, on the other hand, the more explicit information proves more beneficial under these conditions, then the inference would be that greater and more direct help is superior.

To argue that understanding how to solve a problem and understanding the principle underlying the solution may be independent does not imply that both understandings may not be acquired and integrated. Given a limited period of time in which to learn to solve a type of problem, the S with superior intellectual ability would be more likely to accomplish such an integration.

In summary, then, the desirability of making available lesser as against greater

amounts of information in the guidance of problem solving has been a disputed question. Many of the experiments which have been interpreted as arguing for the efficacy of lesser amounts of information are more correctly interpreted as comparing methods involving problem solving as against methods of routine drill. There are other experiments, however, where all Ss were engaged in the discovery of the solution, and where lesser amounts of information proved more beneficial than did greater amounts. One possible interpretation of these results is that there is an advantage accruing to the S when the opportunity for search is not restricted. An alternative explanation has been suggested which would attribute the failure of the more explicit information to the inappropriateness of the kind of information presented.

III. AN EXPERIMENT

Hypotheses

To test the effectiveness of varying *amounts* and kinds of information as guidance in problem solving and with due regard for the argument above, a number of hypotheses were formulated.

It was predicted that only information appropriate to the criterion employed to evaluate success would be effective as guidance. Specifically, where the criterion was the Ss' success in solving problems, both instructional and transfer, only information about the method of attacking specific examples would prove effective. Conversely, where the criterion was success in verbalizing a principle, only information about the principle would prove helpful. In addition, it was hypothesized that, where the information was inappropriate to the criterion

employed, no significant effects would be discerned as a result of the guidance.

It was further hypothesized that, given an appropriate criterion, the effectiveness of the guidance would increase directly as the amount of information supplied the S increased. That is, where the solution of problems was the criterion measure, the more information given about the method of attacking instructional examples, the more successful the Ss would be. Similarly, where stating the rule was the criterion, the more information about the rule which was given as guidance, the more often the principle would be correctly generalized.

Finally, it was hypothesized that the effects of the varying kinds and amounts of information would be similar for Ss of higher and lower intellectual ability.

The Instructional Tasks

To test these hypotheses a problem was required which met certain conditions: (a) the problem had to be one with which most Ss would not have had prior experience; (b) the problem had to be susceptible to numerous variations; and (c) all of the variations had to be solvable by the application of a common principle. The "match-task" described by Katona (9) met these three conditions. An example of this problem-type is shown in Fig. 1.

The directions which accompany the problem shown in Fig. 1 are: "The lines in the drawing represent matches. They have been arranged so that there are five squares in the drawing. Can you make four squares out of the five by moving no more than three matches?"

There are four restrictions which must be observed in solving the problem: (a) all of the squares in the answer must be of equal size; (b) all of the squares in the

answer must be closed; (c) all of the matches must be used in the answer; and (d) all of the sides of the squares in the answer must be only a single match in thickness.

It will be noted that there are 16 matches in the example. Since four squares are required, it follows that no match can serve as the side of more than

ing by the removal of the least number of matches. Knowledge of the principle presumably would help the student to identify these "critical" squares, but it is possible to solve the tasks without becoming aware either of the number of matches in the problem, or of the fact that solution depends on eliminating matches serving a double function. The

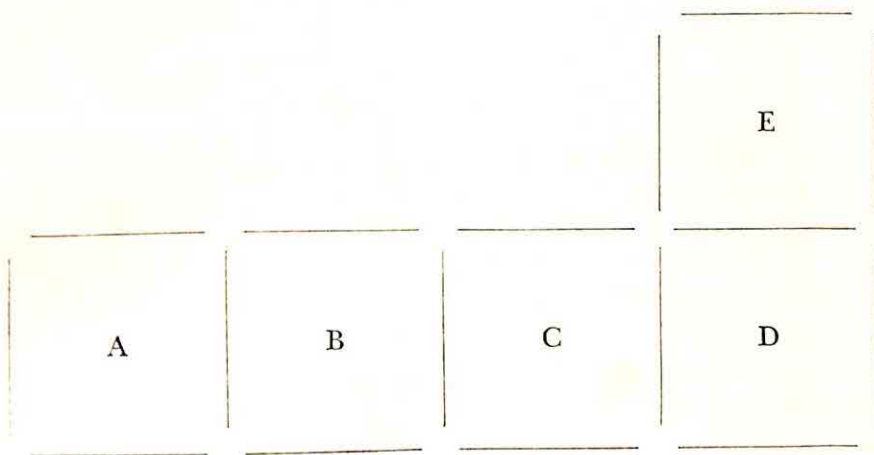


FIG. 1. Sample of Katona match task.

one square in the answer. Matches must be moved, therefore, so that no match serves such a double function. This is the principle common to all variations used in the present experiment.

The principle is not a formula in the sense that knowledge of it will give the S an automatic understanding of which matches are to be transposed in any given variation of the task. Moreover, as Katona points out, solution of the task may be approached quite independently of knowledge of this rule. Solution of the problems may result from perceiving "holes" in the presentation figures. Since the number of matches which one is permitted to transpose is limited, successful solution will hinge on finding those squares which can be completely eliminated from the presentation draw-

ing by the removal of the least number of matches. Knowledge of the principle presumably would help the student to identify these "critical" squares, but it is possible to solve the tasks without becoming aware either of the number of matches in the problem, or of the fact that solution depends on eliminating matches serving a double function. The

rule does serve to limit the alternatives before the S, and should give the problem solver an important clue as to the appearance of a correct answer. The task shown in Fig. 1 was used as a pretest to screen out Ss with prior knowledge of the problem. The Ss were also asked to indicate whether they had ever seen the problem, or ones like it. Those who responded in the affirmative, and who solved the trial problem, were eliminated from the experiment.

Three tasks were used as instructional problems. The first of these was a simple inversion of the puzzle shown in Fig. 1; the square E of that drawing being presented below, rather than above, square D. The second of the instructional tasks represented only a slight modification; square E of Fig. 1 was shifted to a posi-

tion above square B. The last of the three instructional tasks was a somewhat more difficult drawing; square A of Fig. 1 was shifted to a position above square B so that the five squares of the presentation drawing formed a U.

The Instructional Variations

The amount of information given as instruction to the discovery of the rule was varied as follows:

No information. No information was given about the rule, nor was any indication given that a common principle existed.

Some information. The Ss were told that a rule existed, were asked to look for this rule, and the fact that there were 16 matches in the presentation drawing was called to their attention.

Much information. The Ss were told that a rule existed, and were told the rule: "Matches should be moved so that no square has a common side with another square. Every match should be the side of only one square."

The amount of information given as guidance to the discovery of the method of solving the examples was also varied in three ways:

No information. No information was given.

Some information. The Ss were directed to shade in certain of the squares, and their attention was called to the fact that four matches remained which were the sides of unshaded squares. The clue has the effect of calling attention to the critical squares.

Much information. For the three tasks used for instruction the Ss were told which matches to move where, and were asked to practice this solution.

The clues called "some information" were similar, both in the context of a search for the rule and for the method, in that the Ss' attention was called to the structural relationships inherent in the tasks. The clues designated "much information" made available highly explicit guidance and reduced the necessity for an intensive search of the problem by the S. At no point, however, was a completed solution of any task shown to any

S. Every student had to discover a correct response, but this discovery was with the aid of more or less direct help from the instructions.

In combining information about rule and method, the following pattern was adhered to: (a) a statement of the task to be solved, (b) information about the rule, if any, and (c) information about the method, if any. Instructions followed this pattern on each of the three problems on which guidance was given.

Combination of information given about the rule and about the method of solving examples produced nine instructional variations, as follows:

1. None on rule, none on method.
2. None on rule, some on method.
3. None on rule, much on method.
4. Some on rule, none on method.
5. Some on rule, some on method.
6. Some on rule, much on method.
7. Much on rule, none on method.
8. Much on rule, some on method.
9. Much on rule, much on method.

Thus, the instructional variations represented a systematic increase in the amount of information made available to the S from a treatment in which the students were given no information at all to one in which highly directive information was given about both the rule and the method.

The trial problem, a page illustrating and explaining the restrictions, and the three instructional problems (each on a separate page) were developed into an Instruction Booklet to facilitate group administration of the experiment (3). There were nine variations in the practice booklets, each representing one of the alternative treatments described above. The combination of information on rule and method that any particular S received was the same for each of the three instructional problems the S received. The presentation of each task

was accompanied by a space in which the *S* could make provisional tries, and also an answer column in which he could indicate the matches to be transposed in each problem, and could make a drawing of the answer if it had been discovered.

The Test Problems

Eight additional variations of the match task were formulated as test problems. The first three of these were simple reversals of the tasks used for instruction. The final five tasks differed either in that triangles were substituted for squares in the presentation drawings, or in that the problems were made more complex than those used for instruction. For example, Test Task 6 gave the *S* seven squares which had to be changed to five by moving no more than three matches. Test Task 7 presented five squares in the form of a Swiss cross; the compactness of the drawing seemingly creating special difficulties.

Thus, two measures of transfer were provided. These measures will be identified as "simple transfer" for the three reversal test problems, and "complex transfer" for the remaining test tasks.

Finally, a page was given the *S* on which to write a verbalization of the principle common to all variations of the match task which he had encountered.

The test problems and the final page for writing the rule were brought together in a Test Booklet, completely independent of the various Practice Booklets (3).

The Subjects

The materials were administered to 255 twelfth-year students whose median age was 17 years, 3 months. Five students were eliminated on the basis of

prior acquaintance with the tasks. An additional 17 students were absent when the measure of mental ability was secured and were not included. The remaining 233 students represented ten classes in eight high schools in New York and New Jersey suburban communities. A wide variation existed in the facilities provided by these eight schools. Based on information made available by the Metropolitan School Study Council, Teachers College, it was determined that expenditure per pupil ranged from as low as \$165 to \$495. It seems a tenable assumption that the *Ss* had had widely different experiences and types of guidance in problem-solving situations.

Procedures

The two booklets were administered to the *Ss* as follows:

Step 1: Introduction (three minutes). The instruction booklets were distributed randomly within each class. After the students filled in background data, an explanation of the purpose of the experiment was read. The *Ss* were told that neither their intelligence nor ability was being tested and that the results of the experiment would not affect their school grades, but that the experimenter was seeking information on how students learned to solve problems.

Step 2: Trial problem (one minute). Students were asked to attempt the trial problem. At the end of one minute, students were asked to write "Yes" if they had ever seen the specific task or problems like it before; "No" if the problem was new to them.

Step 3: Explanation of restrictions (two minutes). The restrictions were read aloud. Students were then asked to study illustrations of the improper solutions. At the end of two minutes, they were informed that they could look back at the page of restrictions during the remainder of the practice period. This statement was repeated twice during the instruction period.

Step 4: Instruction period (twelve minutes). The students were told that they would have twelve minutes, that they could work at their own best speed, and that they could divide the instruction period between the three tasks as they thought best. Finally, they were informed that since all of the instruction booklets were different, they would not be doing the same thing as other students around them. At the end

of the twelve minutes all instruction booklets were collected. No questions were answered during the instruction period.

Step 5: Test period (sixteen minutes). After the test booklets were distributed, students were informed that they would have sixteen minutes to solve as many of the eight test problems as they could. Again, they were told to work at their own best speed, and to divide the test period among the problems to suit themselves. No questions were answered.

Step 6: Writing the rule (three minutes). At the conclusion of the test period, students were told that they would have three minutes to formulate a rule. They were instructed that they might look back at the test problems they had attempted if this would help them formulate a rule.

With the exception of the trial problem, no attempt was made to assure equal time being spent on each of the tasks. Rather, only the total time assigned to the instruction period, to the test, and to the writing of the rule, was controlled. Experience in preliminary tryouts of the material indicated that withdrawing incomplete tasks was disturbing. Many students abandoned the problems after several had been removed in this fashion. The arbitrary decision to control the total block of time, rather than time spent on each problem, while creating some additional difficulties in the analysis, seemed preferable since a major source of frustration for many students was reduced. More importantly, to have withdrawn the tasks would have penalized most those groups given the least amount of information as instruction.

Test of Mental Ability

All Ss were given Form A of the Otis Self-Administering Test of Mental Ability, Higher Examination (16). The test publishers report a reliability of .92, using comparable forms. The twenty-minute score on this test was employed as an indicator of the Ss' mental ability.

This measure is best thought of as a power score, involving both speed and level of ability (1). The use of time limits during both the instruction and test periods argued for the use of this score as the control of initial ability rather than the untimed Otis score.

Students scoring 45 or better on the Otis were designated as "Highs," or higher than average in mental ability (the mean score for all Ss was 45.3) and those scoring 44 or lower were called "Lows," or lower than average in mental ability.

IV. RESULTS OF THE EXPERIMENT

Number of Students

The number of students receiving each of the nine variations in instructions on the match task was not exactly equal, as shown in Table 1. However, when tested against the marginal totals, the number of students given each form of instruction did not prove to be disproportionate. For this reason it was

TABLE 1
NUMBER OF STUDENTS RECEIVING VARYING
KINDS AND AMOUNTS OF INFORMATION
AS INSTRUCTION

Information about the Rule		Information about the Method			Totals Rule
		None	Some	Much	
Much	Highs	12	15	11	38
	Lows	16	10	14	40
Some	Highs	18	14	13	45
	Lows	8	12	16	36
None	Highs	13	12	14	39
	Lows	11	11	13	35
Totals Method	Highs	43	41	38	122
	Lows	35	33	43	111

Tests for hypothesis that cell frequencies are proportional: Highs: $\chi^2 = 2.5$; Lows: $\chi^2 = 3.9$; $\chi^2 .05, 4 df = 9.5$.

possible to make comparisons of the effectiveness of the varying amounts of information with assurance that the comparisons would not be unduly affected by the existence of unequal numbers in each of the treatment groups.

Mental Ability

An analysis of variance of the mean mental ability scores for the nine treatment means did not provide a basis for rejecting the null hypothesis for either the more or less capable students. In the comparisons which follow, therefore, it was possible to infer that mental ability had been equated for the groups receiving each of the instructional variations.

Instruction Problems

The students were given twelve minutes to attempt the three instruction problems. Since the time the students spent on each of these problems was not controlled, the possibility existed that the students would attempt different numbers of the problems, depending on the kind and amount of information made available to them as guidance. To test this possibility, the mean number of instruction problems attempted was computed for each of the variations. The "number of problems attempted" was defined as the number of the last of the three problems on which some evidence was present that the task had been considered by the student (sketch lines, doodling, etc.).

The analysis of the differences among these means was accomplished by the application of distribution-free tests of significance. These tests were employed to avoid making untenable assumptions both about the normality of the underlying score distributions and the homo-

geneity of treatment variances. These procedures, as adapted, were used throughout the remaining analysis, and are discussed in Appendix A.

When these tests were applied, significant differences in the number of instruction problems attempted were found to exist, attributable to the amount of information the student received about the method. Among the students of higher ability, as the explicitness of the information offered about the method increased, the number of instruction problems attempted also increased. Among the students of lower than average ability, however, while the students who received "some" or "much" information attempted more tasks than did students who received no information, the difference between those for whom the form of help was "some" rather than "much" was not reliable.

No significant differences in the number of instruction problems attempted could be discerned arising from variation in the amount of information the students had received about the rule. Nor did the fact of having received a certain combination of rule and method information result in significant differences in mean number of instruction problems attempted.

To take account of these differences in the number of instruction problems attempted, in analyzing success in solving the instruction tasks, the percentage of those problems attempted which were solved correctly was computed for each of the treatment groups. These percentages are given in Table 2.

The instruction problems (and the transfer problems as well) were scored as either right or wrong. To be credited with having solved the problem, the student had to indicate both a correct

TABLE 2

PERCENTAGE OF INSTRUCTION PROBLEMS ATTEMPTED WHICH WERE SOLVED BY STUDENTS RECEIVING VARYING KINDS AND AMOUNTS OF INFORMATION AS INSTRUCTION

Information About the Rule		Information About the Method			Totals Rule
		None	Some	Much	
Much	Highs	21	58	90	59
	Lows	13	55	65	44
Some	Highs	23	53	90	57
	Lows	00	47	45	40
None	Highs	39	74	86	69
	Lows	09	50	77	51
Totals	Highs	28	61	88	61
	Lows	10	51	62	42

Differences attributable to kind of information received

Kind of information	df	χ^2		
		Highs	Lows	.05
Method information	2	18.0	14.4	6.0
Rule information	2	1.7	8.4	6.0
Method \times rule information	4	4.0	2.8	9.5

Differences attributable to amount of information received

Comparison	U Test			
	Method		Rule	.05
	Highs	Lows	Lows	
Much—no information	3.6	3.6	.0	2.0
Much—some information	2.5	1.4	.4	2.0
Some—no information	2.5	3.3	-1.1	2.0

set of matches to be transposed, and had to make a drawing of the required figure corresponding to the matches marked for transposition.

There were instances in which the student showed a satisfactory answer drawing without marking a correct set of matches to be moved, and the opposite

sometimes occurred. In the majority of cases, however, the two parts of the answer were either both correct or both incorrect. This dependence argued against the use of partial scores. The use of a simple right or wrong score introduced less chance error, since it was not necessary to determine whether the movement of the "critical" matches represented an accidental event or indicated some partial understanding of the task.

For students of higher than average mental ability, the percentage of instruction problems solved increased as the amount of information made available to them about the method was increased. Students given no information about the method attempted and solved fewer of the tasks than students given some information, and these, in turn, attempted and solved fewer problems than students given the greatest amount of information about the method.

For students of less than average ability, differences in the percentage of the instructional tasks solved also accompanied the proffering of information about the method of attacking the examples. For these students, however, no one-to-one correspondence was established between the amount of information presented about the method and the percentage of tasks solved. Giving the less able students information resulted in an increased number of instruction problems attempted and solved, but making available highly directive information did not prove to be substantially superior to giving less explicit clues.

For both the more and the less able students, making available information about the rule underlying the tasks did not result in significant differences in the percentage of the problems solved. If anything, information about the rule re-

sulted in fewer problems being correctly solved. Among students both above and below average in ability, the groups given no intimation that a rule existed, and not told the rule, solved more of the problems than did those who received this information, although the differences were not reliable. The detrimental effects of information about the rule appeared to be somewhat more marked among the less able students.

No significant effects were found in the percentage of instructional tasks solved which could be assigned to the students having been given a particular combination of information about the rule and method.

Simple Transfer Tasks

The first three test problems were simple reversals of the tasks which had been used for instruction. All of the reversals were attempted by all but seven of the students, and these seven were distributed among the treatment groups. The three problems provided a measure of "simple" transfer, since a direct application of the solutions discovered during the instructional period was possible. The mean number of simple transfer problems solved correctly is reported in Table 3.

Differences in the number of simple transfer problems solved were found which were attributable to the amount of information about method which the student had received. For both students above and below average in ability, the effect of having been given some information about the method was to increase the number of simple transfer problems solved. For the more capable students the number of problems correct increased as the amount of information which had been made available in-

TABLE 3
MEAN NUMBER OF SIMPLE TRANSFER PROBLEMS
SOLVED BY STUDENTS RECEIVING VARYING
KINDS AND AMOUNTS OF
INFORMATION

Information About the Rule		Information About the Method			Totals Rule
		None	Some	Much	
Much	Highs	1.5	2.5	2.0	2.0
	Lows	.4	1.4	.9	.9
Some	Highs	.9	1.5	2.6	1.7
	Lows	.0	1.0	.6	.5
None	Highs	1.6	1.7	1.9	1.7
	Lows	.5	1.3	1.7	1.2
Totals	Highs	1.3	1.9	2.2	1.8
	Lows	.3	1.3	1.0	.9

Differences attributable to kind of information received

Kind of information	df	χ^2		
		Highs	Lows	.05
Method information	2	8.4	14.0	6.0
Rule information	2	1.2	9.7	6.0
Method \times rule information	4	10.7	9.4	9.5

Differences attributable to amount of information received

Comparison	U Test			
	Method		Rule	.05
	Highs	Lows	Lows	
Much—no information	2.1	2.4	—1.1	2.0
Much—some information	.9	—1.0	1.2	2.0
Some—no information	1.8	3.4	—1.8	2.0
Much + some—no information	2.2	3.4	—1.7	2.0

creased, although only the difference between the group of students given the most explicit form of information and the group given no information was reliable. For the less able students, on the other hand, the students who had been given the less direct clues solved most of

the problems, although the differences between the two groups given some form of help were not significant. Thus, the high ability students were helped most by the more direct information, while the less able students did not find such help more effective.

Having had varying amounts of information made available about the rule did not result in any substantial differences in simple transfer. While the differences attributable to the three amounts of information given about the rule were not in themselves reliable, the order of performance is of some interest. Among the high ability students, those given a statement of the rule solved more of the simple transfer problems than those not given the rule. Among the less able students, those *not* given the rule or a clue to the rule, solved the most problems.

For the students of less than average ability there was no significant effect resulting from having been given a particular combination of information about both rule and method. Such an effect was present in the performance of the more able students. To evaluate this interaction between the two kinds of information among the more able students, the percentage of correct solutions was determined for each treatment group.

Compared to the group of students who had received no information either about the rule or the method, the two treatments in which the student received an explicit clue in combination with a less explicit clue were significantly superior. Thus, being told the rule but receiving only a clue to the method, or being given directive information about the method with only a clue to the rule appeared to be highly

effective forms of guidance for the more able students. None of the other combinations of information about both rule and method resulted in a performance on the simple transfer problems substantially different from the treatment in which the student received no guidance at all.

The first three test problems were very directly modeled on the problems used for instruction, and it has been shown that there were significant differences in the number of instruction problems attempted and solved. The question arose, therefore, as to what the effects of the treatments would have been on success with the simple transfer tasks if all students had solved a similar number of instruction problems.

Since success on both the instruction and simple transfer problems was affected by the treatments, to adjust the transfer scores for the instruction problem scores would have resulted in removing part of the treatment effects. But if an adjustment were made, and there still remained significant differences attributable to the treatments, then the interpretation would be that not all of the differences on the transfer problems are explainable in terms of the number of instruction problems attempted and solved (2, p. 82f). The inference would be that there were effects arising from the variations in the amount and kind of information quite apart from the percentage of problems solved during the instruction period.

The adjustment was appropriate only for the students of higher than average ability, since only for this group could the hypothesis of no linear regression between the two measures be rejected.

No reliable differences in the means of the treatment groups remained once

the simple transfer scores had been adjusted for percentage of instruction problems solved. It may be inferred that, for the students of greater ability at least, there were no effects resulting from having been given varying kinds and amounts of information as guidance, apart from those which could be explained by the fact that differences existed in the number of instruction problems attempted and solved.

Complex Transfer Tasks

The final five problems of the test departed in form, though the principle underlying the solution remained the same, from the tasks which had been used for instruction. These problems permitted a second measure of transfer. Since these problems were less obvious variations of the instruction tasks, presumably the students whose initial experience with the match tasks had entailed the greatest search should have had some advantage on these problems.

By withdrawing the final task from the analysis, differences in the number of complex transfer problems attempted by the students in the various treatment groups were removed. On the four remaining tasks no significant differences were found in the number of problems attempted. (Using the subgroups described in Appendix A as replications, chi squares of 13.8 for "Highs" and 10.6 for "Lows" were obtained with critical value for chi square .05 equal to 15.5.) Means for the treatment groups on these four problems are given in Table 4.

For the more capable students neither the variations in the amount of information about the rule nor about the method produced significant differences in the number of complex transfer tasks

TABLE 4
MEAN NUMBER OF COMPLEX TRANSFER PROBLEMS SOLVED BY STUDENTS RECEIVING VARYING KINDS AND AMOUNTS OF INFORMATION AS INSTRUCTION

Information About the Rule		Information About the Method			Totals Rule
		None	Some	Much	
Much	Highs	1.5	2.5	2.2	2.1
	Lows	.7	1.0	.7	.8
Some	Highs	1.1	1.3	2.7	1.7
	Lows	.6	1.1	.4	.7
None	Highs	2.0	2.3	1.3	1.8
	Lows	.4	1.7	1.4	1.2
Totals Method	Highs	1.5	2.0	2.1	1.9
	Lows	.6	1.3	.8	.9

Differences attributable to kinds of information received

Kind of information	df	χ^2		
		Highs	Lows	.05
Method information	2	6.9	11.7	6.0
Rule information	2	1.1	4.4	6.0
Method \times rule information	4	10.7	10.7	9.5

Differences attributable to amount of method information received

Comparison	U Test		
	Highs	Lows	.05
Much—no information	1.5	.6	2.0
Much—some information	—0.5	—1.6	2.0
Some—no information	1.6	2.3	2.0
Much + some—no information	1.7	1.6	2.0

solved. Students given information about the method did solve more problems than students given no such information, but the differences in performance were not reliable.

Among the less able students, however, the students who had had the less explicit form of information about the method were successful in solving more of these tasks than were students given

no information about the method. The students given highly directive clues to the method did not differ significantly in performance on these problems from students given less explicit aid or those given no information at all.

Again, no substantial differences in performance on the complex transfer problems resulted from the student having been given varying amounts of information about the rule.

Among both the students of more and less than average ability the groups given no instruction about the method did better relatively, on the complex transfer problems than they had on either the simple transfer problems or on the instructional tasks. As a result, for these four complex problems it was not possible to conclude that the students given information about the method, either in the more or less explicit form, were superior to those given no information of this kind.

For both ability-classes of students, differences were found on the complex tasks resulting from the combination of information received as guidance about the rule and method.

Using the students who had received no information about either rule or method as a comparison group, two combinations of instruction were found to differ significantly among the more capable students. These were the group of students who had been given only a clue to the rule, and the group given only directive information about the method. Both of these groups solved substantially fewer of the complex problems than did the students given no help at all. Solving the greatest number of the tasks were the groups given direct information about the rule together with less explicit information about the method and vice

versa, but these students were not so superior to those given no information at all that chance could be excluded as an explanation of the difference.

On the other hand, among the students of less than average ability, the groups given only a clue to the method, either in the direct or the less explicit form, and given no information about the rule, were superior. None of the other groups of students did significantly differently than the group given no information at all.

Thus, there was little relationship between the order of treatments among the high and low ability students. The students of lesser ability profited most when guidance consisted of information about the method alone, and when no mention was made of the rule. The better students, however, found these forms of help less effective, and were benefited by having received information about the rule in addition to information about the method.

Success on the simple and complex transfer problems was related and the hypothesis of no linear relationship could be rejected. The complex problem scores were adjusted for their regression on the simple transfer scores and the resulting means were compared. No reliable differences in the number of complex problems solved by the various treatment groups remained once this adjustment had been made. Since the adjustment removed part of the treatment effects, the failure to find differences may best be interpreted as indicating that no effects, beyond those explained by the differences in the number of simple transfer problems solved, were present. Thus, success on the more difficult transfer problems was explained by prior success on the simpler tasks. And,

as has been shown for the better students at least, success on the simpler transfer tasks was associated with the students having solved a greater number of the instructional problems.

Writing the Rule

As their final task, students were allowed three minutes in which to write a statement of the principle involved in the problems they had attempted. They were permitted to look back over their work on the match tasks to help formulate this rule.

A statement of the rule was taken to be correct if it referred to the fact that the double function of matches in the presentation drawings had to be eliminated. This principle was sometimes expressed in terms of the statement used in the directions, "No match should be the side of more than one square." But any formulation implying "separateness" was accepted. Thus, "The squares in the answer have to be spread out" was also counted as correct. Between these two degrees of preciseness in verbalizing the rule many variations of expression were possible.

Only 21 per cent of the 233 students in the experiment were successful in correctly verbalizing the rule by this criterion. And since only 12 of the students of lower-than-average ability succeeded in so doing, it seemed inappropriate to analyze the results separately for the two classes of students. The percentage of all students in each of the nine treatment groups correctly stating the rule is given in Table 5.

When the measure of effectiveness of the information given as guidance during the instruction period was ability to verbalize the principle, only the amount of information given about the

TABLE 5
PERCENTAGE OF STUDENTS CORRECTLY
VERBALIZING PRINCIPLE OF MATCH
TASKS

Information About the Rule	Information About the Method			Totals Rule
	None	Some	Much	
Much	39	28	32	33
Some	11	11	28	17
None	29	04	07	13
Totals Method	27	15	22	21

Differences attributable to kind of information received

Kind of information	df	χ^2	$\chi^2 .05$
Method information	2	3.7	6.0
Rule information	2	12.4	6.0
Method \times rule information	4	6.7	9.5

Differences attributable to amount of information about the rule

Comparison	U	U.05
Much—no information	2.5	2.0
Much—some information	.6	2.0
Some—no information	.1	2.0

rule resulted in significant differences.

Neither the information which had been given as guidance to the method, nor the rule and method combinations produced reliably different levels of achievement.

The students who had had the most specific statement of the rule more often stated the rule correctly than those given a clue to the rule or given no information about the rule at all. Those given a clue to the rule were able to verbalize it no more often than those given no information at all.

While the amount of information given about the method did not make for significant differences in successful statements of the rule, the order of the three amounts of such information is

of some interest. Those students who receive no information about the method were most often able to state the rule, followed by those for whom highly directive method information had been the form of guidance. The students given only a clue to the method stated the rule least often. This order parallels that for information about the rule, where the criterion was success in solving the problems, for all students on the instruction problems, and for the "lows" on the simple and complex transfer problems.

A correct statement of the rule was not dependent on the student's prior success in solving the problems. For the total student group (both "lows" and "highs" combined) the following correlation coefficients were obtained:

Success on rule and instruction problems0
Success on rule and simple transfer problems	-.1
Success on rule and complex transfer problems	-.3

None of these correlations was significantly different from zero. (Correlations were between corresponding measures for total subgroups as described in Appendix A. The critical value of the coefficient equals .4.)

V. INTERPRETATION OF RESULTS

A number of experiments have been reported in which seemingly less explicit clues proved more effective as guidance in problem solving than did more direct information, as, for example, information about the principle basic to solution. The superiority of the less explicit information in these situations has been attributed to the more active search demanded of the student, and to the possibility that, as a result of his more intensive search, the student would become cognizant of the structural relations obtaining to the problems.

An alternative explanation was offered which suggested that these experimental findings arose from a misdirection of the search. It was argued that information about principles and informa-

tion about a method of attacking specific examples of a problem type may be two kinds of information, relating to two different subtasks. Guidance given to the discovery of the solution of one of these subtasks, it was suggested, might be inappropriate or even detrimental to the solution of the other, at least in the early stages of solving a problem.

It was predicted that where the criterion of success was appropriate to the kind of information which had been given as guidance, significant effects attributable to the proffering of that information would be discerned, but that where the information given and criterion were inappropriate, no such effects would be found. These predictions held for the present data. Only where the information given was consistent with the criterion used to determine success were significant differences revealed. On the other hand, information not in harmony with the criterion used proved no more efficient than no information at all.

The Ss were apparently presented with two tasks rather than one: (a) to discover a method of attack on the problems, and (b) to discover a satisfactory generalization of the principle. The kind of information given served to direct the search to one rather than the other of these tasks. And, since the more usual pattern in solving problems is first to discover a method of attacking the problem, information given about the principle would seem to introduce special difficulties for the problem solver. In any case, an active search may be assumed to have occurred whatever the kind of information made available to the student as guidance. The direction and focus of the student's search may have varied, however, depending on the kind of information he was given.

It was further hypothesized that there would be no significant difference in the effects of the two kinds of information for students of high and low intellectual ability. This null hypothesis was not substantiated. In the solution of both the instruction and transfer problems (simple and complex), the detrimental effects of having received an inappropriate kind of information appeared somewhat more marked for students of less than average ability. On the other hand, on the transfer problems, the differences attributable to having received an appropriate kind of information were more marked for these less able students. Apparently, for students of lesser ability, the power of information to direct or misdirect was greater than it was for the more able students.

This difference in the power of the information proffered to direct or misdirect may be explained as reflecting a greater ability of the more able students to integrate the two kinds of in-

formation and apply it to the solution of the problems. While no significant interaction was discerned either in the solution of the instruction problems or in the verbalization of the rule, such effects were present on the transfer problems. On the simple transfer problems, the students of greatest mental ability who had received a combination of information about the rule and method were superior, and on the complex transfer problems this superiority was maintained, although on these problems the differences found were not reliable. On the other hand, for the less able students, interaction was found significant only on the more complex tasks, and with these students it was precisely those who had received only information about the method, and no information about the rule, who were superior.

To suggest that information about principles and about methods of attacking a problem are two kinds of information does not imply that a combination of the two kinds of information cannot be understood and applied in problem solving. To derive a principle from an understanding of method, or to derive a method of attacking problems from an understanding of principle, and to integrate and apply the two understandings, would require more time than to discover either method or principle alone. Where the time allowed for problem solving is limited, students of highest mental ability might be expected to accomplish this integration more often than students of lesser ability. Given enough time, relatively easier examples, and greater success in solving the easier tasks, students of less-than-average ability might also be expected to benefit from receiving both kinds of information. But apparently these conditions did not obtain for the poorer students in their work with the match tasks.

It may be concluded that while information about principle and about method are two kinds of information, the effects of having received one rather than the other kind are more likely to be noted in the early stages of learning to solve a class of problems, and more likely to appear among poorer than among more able students. (And, though the experiment does not provide for this, it might be hypothesized that the effects are more likely to appear as the difficulty of the problem to be solved is increased.)

If information about method and about rule are accepted as two kinds of information, the test of the effectiveness of greater as against lesser amounts of information as guidance must be made where the criterion used to measure effectiveness is appropriate to the kind of information supplied. It was predicted that where an appropriate criterion was employed, the effectiveness of guidance would increase directly as

the amount of information supplied the *S* increased. It was hypothesized that this relationship would hold both for students of greater and lesser mental ability.

Where the criterion was success in solving instructional and transfer problems, the predicted relationship between the amount of method information received as guidance and success was more characteristic of the students of greater ability, though it was not fully substantiated even among these students. On the instruction problems, it was true that the more method information the able students were supplied, the greater the percentage of the problems attempted and solved. This direct one-to-one relationship began to weaken on the simple transfer problems. On these tasks students who received information on the method were superior to those who had received no information, but the difference between the students who had received the greatest amount and those who had been given less explicit clues was not reliable. On the more complex transfer tasks, no reliable differences could be noted between the *Ss* receiving the three different amounts of information. Thus, while there appeared to be an initial advantage accruing to those receiving the greater amount of method information, as the more able students tried more of the problems, this initial advantage did not persist.

For the less able students advantages resulting from having been given highly directive information about method were even less marked. For these students, not only did the predicted relationship between amount of information about method and success in solving examples fail to appear, but, in fact, some evidence was secured to demonstrate that the less explicit form of guidance was preferable. On the complex transfer tasks, for example, only the *Ss* receiving the less explicit form of information were superior to those receiving no information at all. And since the more able students receiving the greatest amount of information as instruction failed to maintain their initial advantage, there is some support for the advocacy of those methods of guidance which require greater search and which lead to greater awareness of structural relationships.

But this interpretation is not entirely satisfactory. In the first place, the "awareness" or "structural" explanation does not make it clear why, if less explicit clues were beneficial for the less able students, such methods were not even more beneficial for students of greater than average ability. Second, if having given the students "some" information proved relatively more advantageous on the transfer problems than "much" information, and if this is explained by the greater search required, then why did not

the students receiving no information about the method do best of all? If a little search is a good thing, should not more search have been even a better thing? Third, why did those given the lesser amounts of information about the rule do less well in writing the rule? If search is helpful in and of itself, then similar benefits should have been exhibited on this task as well.

An alternative explanation of the results obtained in this study is possible. On the transfer tasks the principal differences were accounted for by the number of instruction problems attempted and solved correctly. For the complex tasks no reliable differences remained, attributable to the treatment variations, once the differences explained by prior performance on the simpler tasks had been taken into account. On the simple transfer tasks, in turn, no differences remained once variation in the percentage of instruction problems solved had been removed.

The advantage of having been given some guidance to the method was that it enabled the student to attempt and solve more of the initial problems in the limited time allowed. Whether the more or the less explicit form of guidance proved more effective depended upon the ability of the student to understand and apply the information given. The extent to which information was understood and utilized was not solely a function of the amount given. The better students were able to apply directly to the task at hand greater amounts of information. Less able students were less likely to comprehend the implications involved in the greater amounts of information, could less often apply it, and for these students, even on the initial problems, no differences appeared as between those given direct information and those given only a structural clue.

The fact that the more able students could make use of the more direct form of information may also explain the results obtained in stating the rule. On this task, students for whom the rule was identified succeeded more often than those given a clue to the rule, and these successful students were, with but a few exceptions, all drawn from the upper ability group.

Moreover, the differences in interaction for the two groups of students are also explained by the differences in their ability to make use of information given. Among the better students it was those given a combination of rule and method information who did best, but among the poorer students it was precisely the groups given information about the method, and none about the rule, who did best on the transfer tasks.

As the students moved from the instruction problems to the simple transfer tasks and then to the more complex tasks, the differences resulting from their having had a certain type of guidance rather than another became less

marked. This was true both for the more and less able students, but more especially for those of greatest ability. This suggests that the beneficial effects of guidance lie in the speed with which the student is brought to his best level of performance. Understanding, to the level of the S's ability, is likely to be achieved irrespective of the kind and amount of guidance offered. If the guidance is appropriate, and if the student is able to make use of the greater amounts of information, he will be helped to attain his maximum level most quickly by being given more direct forms of aid. But if the direct instructions cannot be applied by the student, less direct forms of instruction may prove just as beneficial.

The present study has dealt with the solution of intellectual problems. On the basis of the results obtained it might be argued, but by analogy only, that in the guidance of emotional problems the efficacy of nondirective guidance may result from the inability of the S to apply specifically directive information to what are, for the Ss, extremely difficult problems.

The ability to verbalize a principle without, at the same time, a corresponding ability to apply the principle to the solution of problems, is not generally accepted as a satisfactory outcome of guidance. With this in mind, the results of the present experiment carry certain implications for guidance in situations similar to those experimented with here.

The learning situation used was characterized by the following conditions: 1, students were engaged in intellectual problem solving—they had to discover the solution for themselves; 2, the problem was one in which previously acquired generalizations could not be readily applied; 3, the principle underlying the problem was not a formula in the sense that it prescribed specific steps needed to solve any given variation of the problem; and 4, the time given to instruction and problem solving was restricted.

In situations meeting these conditions, the results of the study suggest:

1. Information given the student about the method of solving examples is more likely to be beneficial than information given about the principle—at least in the initial stages of problem solving.

2. Some appropriate guidance will prove more helpful to the student than no guidance. Leaving the student to discover for himself the solution of a problem will not prevent understanding, but will probably delay it.

3. The effectiveness of guidance does not depend solely on the amount of information imparted. More explicit forms of instruction will prove most helpful with students able to apply

the information. For students of lesser ability, less explicit clues, designed to highlight structural relationships, may prove just as effective.

VI. SUMMARY

An unresolved issue in the guidance of problem solving is the desirability of making available greater rather than lesser amounts of information. Information may be of two kinds; it may relate either to the principle or to method. It was hypothesized that performance would improve as the amount of information given as guidance about method of solution or principle for solution was increased.

Several forms of the match task were presented to twelfth grade students. In guidance there were three variations in the amount of verbal information given about the principle: "No information"; "Some information," a clue to the principle; and "Much information," a statement of the principle. Similarly, students were given "No," "Some," and "Much" information about the method of solving examples. The combination of the amounts of information about method and principle yielded nine different instruction treatments ranging from one with no information about either the principle or the method, to one with highly directive information about both principle and method. Success was appraised in the number of instructional tasks solved, in transfer to simple and complex problems, and in verbalizing the principle separately for students above and below average in mental ability.

For students above average in mental ability, success in solving problems increased only for the instruction problems directly as the amount of information given about method increased. For students of lower ability, some appropriate guidance was superior to no guidance. In

solving the problems both for students above and below average in mental ability, information about the rule did not seem to affect results differentially.

For the simple transfer problems, the more capable students seemed to profit most from the combination of explicit information about the rule with an indirect clue to the method, or from a clue to the rule in combination with explicit information about method. On the complex transfer tasks, the high ability students receiving only information about the method, or only an indirect clue to the rule, did less well than students given other combinations of information. But on these complex tasks, students of less than average ability given only guidance to the method, in either the direct or less explicit form, were superior.

No treatment effects remained for the complex transfer problems once an adjustment had been made for differences on the simple transfer tasks. Similarly, for the high ability group, at least, no treatment effects remained on the simple transfer problems once scores were adjusted for differences in success on the instruction problems.

Success in writing the rule increased as the amount of information students had been given about the rule increased, although only the students given the most explicit information about the rule were superior to those given no information about the rule. In writing the rule, no effects were noticed which could be attributed to the amount of information the students had received about the method of solving examples. Success in verbalizing the rule was uncorrelated with success in solving the problems.

These results were interpreted to indicate (a) that information used in guidance must be appropriate to the task set for the student, (b) that some appro-

priate guidance is beneficial, but that failure to provide it will delay rather than prevent solution, (c) that the effectiveness of guidance does not depend solely on the amount of information imparted, but that (d) more explicit in-

struction will prove most helpful with the more able students while (e) less explicit instruction may be just as effective as more directive guidance for the less able students.

APPENDIX A

STATISTICAL PROCEDURES

The analyses of variance reported in the text, with the exception of the analysis of mental ability means, were carried through on ranked measures. In this way, the assumptions of normality of score distributions and of homogeneity of treatment variances were avoided—assumptions which in most cases were untenable. It is interesting to note that the practical difference of this use of distribution-free statistics was not great. As a check, all of the analyses were carried through in the more traditional manner in disregard of the assumptions. Except for a few instances in which the nonparametric test statistic proved to be very close to its critical value, the decisions reached were identical under either procedure. However, the distribution-free tests described below were the techniques of choice for the present data.

For each of the nine treatments, and for both the "Highs" and "Lows," three subgroups were formed. This was done by assigning students to the subgroups, in order, from a list of the ranked mental ability scores for each treatment. For example, for the treatment: "No information on rule; no information on method," the subgrouping was as follows:

Otis Score	Subgroup
69	1
60	2
54	3
54	1
54	2
53	3
.	.
46	1

Students in the other treatment groups were assigned to subgroups in a like manner. The composition of the subgroups remained constant for all of the analyses, of course.

The next step was to compute means (or, in some cases, percentages) for each of the subgroups on the particular measure being studied. To test the main effects of the three variations in information about the method, the variations in information about the rule were treated as

replications, with comparable subgroups ranked. To test the main effects of the variations in information received about the rule, the process was reversed, and the comparable subgroups on method were treated as replications. The test itself consisted in determining the probability of the obtained rank totals for, first, the three variations in method information, and then for the three variations in rule information. The formula used and the test is more fully described by Moses (15) and Freidman (7).

To test the significance of the interaction, Rule \times Method, the procedure suggested by Wilcoxon (24) was employed. The test statistic was computed as a sum of two values. One component was obtained by tabulating the differences between corresponding subgroup values for the "Much" and "Some" rule variations for each of three variations in method information. The second component was obtained by finding the difference between the sum of the values of the "Much" and "Some" rule variations and twice the value of the "None" rule variation for comparable subgroups—again, for each of the three method variations. These differences were then ranked and the test described above was applied to the rank totals. The argument for the procedure is that, assuming there is an advantage accruing from having received a particular combination of rule and method information, the differences will be consistently greater (or smaller) for each of the replications of the treatments so devised.

In those cases where the main effects were shown to be significant by the procedure described above, the Mann-Whitney (11) "U" test was used to analyze the difference between individual means of the variations. The test is well known and does not need to be described, but it needs to be pointed out that the individual comparisons were made by ranking the eighteen subgroup means for any two variations under consideration.

Finally, it should be noted that correlations and tests for linearity, which are reported in the text, were also based on the subgroup estimates on the measures of interest.

REFERENCES

1. BAXTER, B. An experimental analysis of the contribution of speed and level in an intelligence test. *J. educ. Psychol.*, 1941, 32, 285-296.
2. COCHRAN, W., & COX, G. *Experimental designs*. New York: Wiley, 1950.
3. CORMAN, B. R. *The effects of varying amounts and kinds of information as guidance in problem-solving*. Unpublished doctor's dissertation, Columbia Univer., 1954.
4. CRAIG, R. *The transfer value of guided learning*. New York: Teachers College, Columbia Univer., 1953.
5. DUNCKER, K. On problem-solving. *Psychol. Monogr.*, 1945, 58, No. 5 (Whole No. 270).
6. DURKIN, H. E. Trial and error, gradual analysis, and sudden reorganization. *Arch. Psychol.*, 1937, No. 210.
7. FRIEDMAN, M. The use of ranks to avoid the assumption of normality. *J. Amer. statist. Ass.*, 1937, 32, 675-701.
8. HILGARD, E., IRVINE, R., & WHIPPLE, J. Rote memorization, understanding, and transfer. *J. exp. Psychol.*, 1953, 46, 288-292.
9. KATONA, G. *Organizing and memorizing*. New York: Columbia Univer. Press, 1940.
10. LUCHINS, A. S. Mechanization in problem solving. *Psychol. Monogr.*, 1942, 54, No. 6. (Whole No. 248).
11. MANN, H. B., & WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 1947, 18, 50-60.
12. MARKS, M. R. Problem solving as a function of the situation. *J. exp. Psychol.*, 1951, 41, 74-80.
13. MCCONNELL, T. R. Discovery versus authoritative identification in the learning of children. *Univer. Iowa Stud. Educ.*, 1934, 9, No. 5, 13-62.
14. MORGAN, J. J. Effect of non-rational factors on inductive reasoning. *J. exp. psychol.*, 1944, 34, 159-168.
15. MOSES, L. Nonparametric statistics for psychological research. *Psychol. Bull.*, 1952, 49, 122-143.
16. OVERMAN, J. R. An experimental study of the method of instruction on transfer of training in arithmetic. *Elem. Sch. J.*, 1930, 31, 183-190.
17. OTIS, A. S. *Otis self-administering tests of mental ability*. Yonkers, New York: World Book Co., 1928.
18. RUGER, H. A. *The psychology of efficiency*. Teachers college education reprints, No. 5. New York: Teachers College, Columbia Univer., 1926.
19. STACEY, C. L. The law of effect in the retained situation with meaningful materials. In Esther Swenson, G. L. Anderson, and C. L. Stacey, *Learning theory in school situations*. Minneapolis: Univer. of Minnesota Press, 1949.
20. STROUD, J. *Psychology of education*. New York: Longmans Green, 1946.
21. THORNDIKE, E. L. *The psychology of wants, interests, and attitudes*. New York: Appleton-Century, 1935.
22. WATERS, R. H. The influence of tuition upon ideational learning. *J. gen. Psychol.*, 1928, 1, 534-547.
23. WERTHEIMER, M. *Productive thinking*. New York: Harper, 1945.
24. WILCOXON, F. *Some rapid approximate statistical procedures*. Stamford, Conn.: Amer. Cyanamid Co., 1949.

(Accepted for publication February 26, 1956)



Psychological Monographs: General and Applied

Failure-Avoidance in Situational Interpretation and Problem Solving¹HAROLD M. SCHRODER
Princeton University

AND

DAVID E. HUNT
Yale University

with the collaboration of JOHN McDAVID, JR., Princeton University

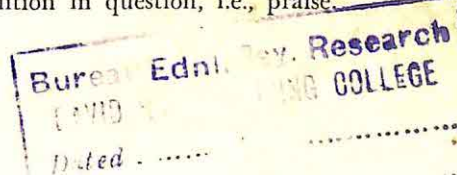
I. THE PROBLEM

THE IMPORTANCE of understanding the relationship between personality factors and problem-solving behavior has recently become apparent (4, 10, 18). The purpose of the present study was to investigate several hypotheses concerning the relationship between personality and problem-solving factors *under negative conditions*. That an individual's reaction to a negative state of affairs in the environment is important becomes apparent when one notes the numerous personality constructs employed to con-

ceptualize such reactions: stress tolerance, frustration tolerance, ego strength, and tolerance for ambiguity. It is implicitly assumed in the use of such constructs that a person's inability to cope with environmental adversity is more or less central to many adjustment problems; such an assumption is explicitly supported in the repeated incidence noted in case histories of a negative condition in the patient's life situation precipitating a breakdown in adjustment. Therefore, we felt justified in our selection of *failure* and *criticism* conditions for study, and have accordingly employed these situational conditions not only as experimentally induced conditions, but also as conditional items in personality measures.² Such a restriction of conditional (or situational) variance represents our agreement with Rotter (19) on the importance of taking account of, and if possible controlling, the psychological situation. It is not unlikely that much of the previous difficulty in accounting for interindividual variance in problem solving by the use of personality instruments

¹This investigation was supported in part by a research grant M-955 from the National Institute of Mental Health of the National Institutes of Health, Public Health Service, and in part by the Rockefeller Foundation for research in perception and social psychology. We are greatly indebted to the staff members of the following participating organizations and institutions for their generous cooperation: YMCA camps with headquarters in Trenton and Newark, N.J.; Ten Mile River Boy Scout Camps (N.Y.); Sequassen Boy Scout Camp (Conn.); Valley Regional High School, Deep River (Conn.); Hamden High School (Conn.); Notre Dame High School, West Haven (Conn.); Fairfield High School (Conn.); Perth Amboy High School (N.J.); Olney High School, Philadelphia (Pa.); The Peddie School, Hightstown (N.J.); and The Hun School, Princeton (N.J.). We would also like to express our appreciation to our research assistants Mr. Ronald S. Wilson and Mr. Arthur Adlerstein.

²A part of one section (V) also noted reactions to praise conditions, but here also the relevant personality items were appropriate to the experimental condition in question, i.e., praise.



(e.g., 8) has been due to the implicit lack of recognition of such situational specificity, as evidenced by the previous use of general measures such as the Rorschach.

This study investigates a number of hypotheses bearing either directly or indirectly on the defensive process of *avoidance*. Other descriptive terms such as denial or blame-projection (17) are closely related to our use of avoidance. In our use of the term avoidance we will mean a *behavior or interpretation directed away from a situational event*. For example, in the situation of failure, a person's quitting a task when he is performing poorly may be considered failure-avoidance *behavior*, while a person's refusal to perceive the possibility of his own inadequacy under these same situational circumstances may be regarded as an *interpretation* directed at avoiding failure.

Section II describes a method for measuring four behavioral variables in a single problem-solving situation under conditions of failure and presents the relationships between these measures. Using primarily speed and error measures, previous studies (e.g., 13) have demonstrated that failure increases interindividual performance variation. Regarding this disruptive effect of failure upon the performance of certain subjects (Ss), Lazarus and Eriksen state, "It is also possible that some Ss did much worse on Test 2 because they became disgusted by the whole situation, or so threatened that their response was to *give up or to 'psychologically leave the field'*" (13, p. 103, italics ours). Since avoidance is a potentially important response to failure, we therefore devised a behavioral method to permit an objective measure of this "leaving the field"

response, along with three other measures.

Section III describes a personality measure intended to assess the probability of a person's making an *avoidant interpretation* to a failure situation, i.e., refusal to perceive that he might be failing. In Section IV the relationship between this failure-avoidant interpretation measure and failure-avoidant behavior is explored. Implicit in our conception of failure-avoidant interpretation is an underlying defensive process which prevents the individual from lowering his evaluation of himself under negative conditions. Sections V and VI therefore investigate the relationship between failure-avoidant interpretation and self-evaluation following experimentally induced negative conditions.

II. RELATIONSHIP OF FAILURE-AVOIDANT BEHAVIOR TO EXPECTANCY, BEHAVIORAL ALTERATION, AND ERRORS IN A PROBLEM-SOLVING SITUATION

This section has two aims: (a) to develop a method for describing an individual's behavior in a problem-solving situation along four behavioral dimensions—expectancy, looking for alternative solutions, quality of performance, and failure-avoidance; and (b) to study the relationship of failure-avoidant behavior to the other three measures. These variables are referred to as behavioral measures because each one can be measured by direct observation. Although the present study used behavioral measures, it is assumed that such data differ from personality test data only in the method employed and not in the nature of the variables involved. After measuring the interindividual variability on several behavioral variables, one can then undertake the problem of accounting for this

variance through the use of personality constructs and derived instruments.

The importance of failure-avoidant behavior has been previously discussed and may be re-emphasized by noting that if an individual leaves or avoids the problem it is obvious that there can be no solution. Therefore, because avoidance occupies this central position in the problem-solving process, all hypotheses were formulated relative to avoidance. The other three variables selected for study have been employed and discussed previously in the literature, though not always in a problem-solving context and not always with the same operations used here. Definitions of the four measures follow:

a. *Failure-avoidant behavior*: extent to which *S* avoids a situation in which he is receiving gradually increasing failure by selecting a different goal rather than continuing to strive toward the same goal.

b. *Expectancy or goal-getting behavior* (20): the score which *S* reports that he actually expects to make on the next trial, i.e., the traditional level-of-aspiration measure.

c. *Behavioral alternation or looking for alternative solutions* (23): the number of times *S* changes his solution or approach while striving toward the same goal. This variable has been previously assumed to be of importance as indicated by the following statement from Eriksen, Lazarus, and Strange: "Giving subjects false information that they are failing may cause some of them to alter their mode of attack on the problem" (8, p. 284).

d. *Quality of performance* (6): relative error measure based on pre- and post-failure performance.

The following hypotheses were tested:

1. That failure-avoidant behavior will

be directly related to expectancy, i.e., *Ss* who set high goals will be more likely to avoid the failure situation. Since high goal-setting behavior has been previously regarded as "unrealistic" (18, 21), we assumed that such behavior therefore represented one form of avoidance.

2. That failure-avoidant behavior will be inversely related to behavioral alternation, i.e., *Ss* who use fewer solutions will be more likely to avoid the failure situation. In a previous study investigating behavioral alternation, Schroder and Rotter (23) have suggested that rigid behavior (i.e., not looking for alternative solutions) may be viewed as one form of avoidance as follows: "The avoidance behavior itself may be regarded as a single solution which is regularly reinforced by preventing the reoccurrence of some trauma, thus providing the conditions making for single solution learning" (23, p. 148).

3. That failure-avoidant behavior will be directly related to performance disruption, i.e., *Ss* who make more errors after failure will be more likely to avoid the failure situation. This hypothesis is based on the assumption that the tendency to make errors following failure stems from the same kind of defensive process which leads to failure-avoidant behavior.

In the method described the problem-solving situation consists of an insoluble sorting problem with six available but incorrect solution pathways to a desired goal. The *S* is free to use as many or as few of the available solutions as he chooses, and is also free to avoid the situation by choosing an alternative goal (i.e., a different problem), or to continue to try to reach the same goal. Progress toward this goal was indicated by controlled scores given by *E* as well

as *S*'s knowledge that his solution was incorrect. In previous studies, failure has frequently been induced either by a combination of referents, e.g., social reproof accompanied by a low score (22), or by very strong failure stimulation (14). The present method utilizes a technique of minimal and gradually increasing failure with increasing trials. An attempt was made to specify the failure stimulation as much as possible so that only the score given to *S* would be the referent for failure. Since the induction of failure inevitably involved an interpersonal situation between *E* and *S*, it became important to insure that the experimental condition and the obtained results could be replicated. Therefore, two separate investigations using the same method were carried out in different localities for purposes of cross validation.

A. Method

1. Subjects

Group A: 57 male freshmen psychology students at University A.

Group B: 78 male freshmen who had not been in a psychology course, secured from the student employment service at University B.

2. Materials

a. A sorting box containing four slots and so constructed that *S* was unable to see a card after it had been sorted.

b. Task 1 was an insoluble card-sorting problem with six theoretically possible solutions. On each trial, a pack of 12 cards was used, each card containing a cue variation of six major concepts: color of card (blue, yellow, pink, and green); position of letter (upright, left, right, and inverted); nonsense syllable (MUV, REB, NUS, and JOH); form (circle, square, triangle, and pentagon); number of forms (from one to four); and color of form (black, gray, white, and red). Each major concept was varied throughout the 12 cards by the four cue variations indicated parenthetically so that one slot in the problem-solving box could be used for each variation in the major concept (solution)

which *S* chose to use. For example, the first card was constructed as follows: blue card (card color); letter inverted; MUV (nonsense syllable); two (number) white (color of form) squares (form). The distribution of four cues within each major concept was always 4:3:3:2 so that it was impossible to achieve the instructed goal of obtaining an equal distribution without making an error. Four different packs were employed and the cards were presented to *S* in a constant order which maximized the distribution of cues over the four slots during the early sorts.

c. Task 2 was present only so that *E* could point to it during the instructions, since the experiment was completed whenever *S* left Task 1. Task 2 consisted of a series of questions typed on sheets of paper and placed in a folder.

d. Sets of norms which *E* used ostensibly to calculate *S*'s score after each trial.

e. Score slips and record sheets, used as described in procedure.

3. Instructions

The experiment was carried out by two *E*s in two different localities using identical instructions and procedure. The instructions³ presented the problem as follows: *Task 1* was structured as a measure of decision-making ability and leadership potential, and *Task 2* was structured as a different task measuring a different ability, creative intelligence. The instructions were designed to induce ego involvement and to emphasize the importance of succeeding by pointing out the relationship between these measures and postcollege success. Then *E* continued as follows:

"Task 1 (pointing) measures the manner in which you handle decision making or, more generally, your leadership potential. It consists of sorting or organizing packs of 12 cards into four groups differing along a single characteristic. The aim is to sort the cards into four groups (pointing to four boxes)." Examples were then given to insure that *S* understood the nature of the problem. Then *E* continued:

"When I tell you to start, you will select a single card, look at it, make your decision about grouping and place it in one of the boxes. Note that you cannot see a card again once it is sorted into one of these sorting boxes (pointing). Then select a second card from the supply box and sort it. Continue selecting cards one at a time and sorting them into one of the four relevant sorting boxes until the 12 cards are used.

³ Detailed instructions for all procedures may be obtained from the authors.

"Your score will then be worked out by me on the basis of the following:

"First, your score will depend upon the distribution of the cards. The more even the distribution of the cards over the four boxes, the higher the score. Of course, the best solution is four groups of three each. The nearer this is approached, the higher the score and the lowest score would be for 12 cards in one box.

"Second, your score will depend upon the time taken, so you should work as fast as you can.

"Third, your score will depend upon the errors made. An error consists of placing a card in the wrong box.

"And, last, your score will depend also upon your general approach to the problem.

"Unlike most other tests you have probably taken, in this task you will be given a good deal of information. This is because it is a test of your decision making. After each trial your score will be worked out by me. Once the test starts, there should be as little talking as possible. In order to facilitate this, I will pass the score you actually made on that trial to you on a slip of paper like this (hand score slip to *S*). At this point you can enter your score on the record sheet provided (hand record sheet to *S*) for your own information. After you have entered your score you have two things to record on this slip.

"First, select the task which you wish to try on the next trial. This choice is up to you. It is perfectly OK if you want to leave Task 1 and try something else after any trial.

"Second, enter the score that you actually expect to make on the next trial. When this is done, pass the sheet back to me and wait for me to give you the signal to start (take slip from *S*).

"Now on this task do not pay any attention to the back of the cards. The symbols there are for identification purposes only and have nothing to do with the task. Select only one card at a time, turn it over so that you can see the front, make sure it is not upside down, then place it in one of the boxes and select another card. Remember to sort on the basis of a single characteristic. When you have sorted all 12 cards, say "finished" so that the exact time can be recorded. The materials may differ from trial to trial, but the problem remains the same. This task is not essentially difficult, but your initial reaction to the task will very likely be that it is a little complex. However, as you go along you will most probably see the basis of the solution. Since your score is based on a number of different factors—distribution, speed, errors, and approach—your score should get higher since improvement may occur on any one

of these factors. That is, you should improve as you go along . . . up to a level of somewhere between 30 and 40 which is the average for college freshmen. Are there any questions? Do as well as you can . . . ready begin."

4. Procedure

After each trial, *E* entered the time taken and the code number of the cards placed in each sorting box. The *E* then referred to the "norms" and entered *S*'s predetermined score on a slip of paper which he passed to *S* (in order to minimize any possible biasing effects of extraneous social factors). The slip contained three statements:

"Your score on trial was

Check *a.* or *b.*

a. Desire to continue with task 1

b. Desire to try task 2

Write down the score you actually expect to make on the next trial

The *S* entered his score on the record sheet provided, checked one of the two tasks, recorded his expectancy, and returned the slip to *E*. The *E* inspected the slip to note the task choice, placed the next pack of cards in position (four packs of cards were rotated throughout the trials) and signalled *S* to begin the next trial. This procedure was continued until *S* selected Task 2 or until the twelfth trial had been reached. In either case, a half-hour interview related to *S*'s attitude toward the problem and attempted solutions was then conducted. Following this interview, *S* was informed of the true nature of the experiment and requested not to discuss the experiment with other students.

5. Score Pattern

The score pattern for consecutive trials was: 9, 13, 16, 20, 21, 22, 22, 21, 20, 18, 17, and 15 (e.g., score on the fifth trial was 21). The early scores were low relative to the instructions and the expectations of college students. The scores advanced to a point still below the minimal goal, leveled off, and then decreased markedly. It was assumed that *S* would react to each score in terms of certain elements in the instructions which indicated that (*a*) *S* should improve with practice, and (*b*) a good score would be 30 or more.

6. Operations and Scoring

The cards were code-numbered so that *E* could identify both the concept (solution) used on a given trial and the number of errors. Objective scoring was possible in all except three or four cases where the solution for a single

trial was doubtful. In these cases, *S*'s report in the interview of the solution used was accepted. An error was defined as a misplaced card in terms of the concept being used for sorting on that trial. Expectancy and avoidance measures were directly available on the score slip filled out by *S*.

Since *S* was free to leave Task 1 at any time after the first trial, it was necessary to make some arbitrary decision regarding the trial an *S* must reach before he would be included in analysis. Trial 5 was selected because this was the first point at which the score pattern began to level off. That is, if *S* left Task 1 before Trial 5 it is likely that he either misunderstood the instructions or placed a higher value on Task 2, and in either instance he should be excluded. Therefore the use of Trial 5 as the cut-off point should provide a good sampling of *S*'s behavior on the variables under consideration.

Of the total of 135 *S*s in both groups, 44 were dropped for the following reasons: 25 because of their selecting Task 2 before Trial 5; 13 because of their previous participation in an earlier experiment involving false scores which led to their being suspicious of scores as indicated in interviews; three because of their color blindness; and three because of their misunderstanding of the instructions and sorting on the basis of the code numbers on the back of the cards.

The following measures were used for each variable: (a) *Failure-avoidant behavior* was measured by trial at which *S* selected Task 2 or left Task 1; since the experiment was discontinued after Trial 12, an avoidance score of 13 indicated that *S* did not leave Task 1 at all (and this would be considered low avoidance).

(b) *Expectancy* was calculated by adding *S*'s expectancy scores for the first four trials. (c) *Behavioral alternation* was measured by counting the number of solution alternations during the first five trials, giving a range of 0 to 4. (d) *Quality of performance* was measured by comparing the errors on the same solution before and after Trial 5; i.e., as failure stimulation increased, did errors increase, remain the same, or decrease? This relative measure was calculated by comparing the number of errors *S* made on a given solution on Trial 6 with the number of errors made on the previous trial in which that solution was employed. If *S* used a new solution on Trial 6, the next trial for which an earlier solution was available for comparison was used (this procedure of comparing relative error-making tendencies only on the same solution was intended to control for the differential difficulty level of the six solutions).

B. Results

Table 1 gives the cutting points and number of *S*s for each of the four variables in Group A, Group B, and combined group. With the exception of the error measure, a pooled distribution of the two groups combined was used to establish cutting points which would yield the most nearly equal number of *S*s in each subgroup. Using this procedure, avoidance and alternating were split

TABLE 1
GROUP AND COMBINED DISTRIBUTIONS FOR BEHAVIORAL MEASURES

Group	Expectancy				Alternating			
	61-72	73-79	80-89	90-102	0-1	2	3-4	N
A	11	12	11	8	15	11	16	42
B	12	13	12	12	21	11	17	49
Total	23	25	23	20	36	22	33	91
	Errors*				Avoidance			
	Decr.	Same	Incr.	N	12-13	9-11	5-8	N
A	11	18	11	40	18	13	11	42
B	13	14	15	42	16	14	19	49
Total	24	32	26	82	34	27	30	91

* The group totals for errors are 40 and 42 from Groups A and B rather than 42 and 49 since error measures do not include *S*s who switched to Task 2 at Trial 5.

TABLE 2
RELATIONSHIP OF EXPECTANCY, ALTERNATING, AND ERRORS TO FAILURE-AVOIDANCE

Hypothesis	Group	χ^2	df	p	ϵ^*	F	df	p
1. Expectancy-avoidance	A	18.002	6	< .01				
	B	11.188	6	< .10				
	Total	29.190	12	< .01	.41	2.02	18/72	< .05
2. Alternating-avoidance	A	8.200	4	< .10				
	B	6.900	4	< .20				
	Total	15.100	8	< .06	.33	3.75	4/86	< .01
3. Errors-avoidance	A	14.820	4	< .01				
	B	6.663	4	< .20				
	Total	21.483	8	< .01				

* The epsilon reported is that of the regression of avoidance on the other variable, e.g., expectancy, since the hypotheses were stated in this form and the measures obtained in that order.

in thirds, and expectancy in fourths. The error measure was split according to the logical procedure of errors decreasing, remaining the same, or increasing as failure stimulation increased since these three categories yielded groups of fairly equal size.

The groupings in Table 1 were used to calculate chi-square values for testing the hypotheses, as summarized in Table 2. In addition, the degree of relationship existing in Hypotheses 1 and 2 was evaluated by the use of epsilon (since the avoidance distribution was obviously nonnormal) computed on the basis of raw scores rather than grouped data. Epsilon was not calculated for Hypothesis 3 since quality of performance was a relative measure.

Hypothesis 1. Avoidant behavior is directly related to expectancy. That is, Ss who set high goals are more likely to avoid the situation while low-goal-setting Ss are more likely to continue to strive toward the same goal under failure conditions.

Hypothesis 2. Avoidant behavior is inversely related to behavioral alternation. Although the relationship as reflected

by chi-square value is at borderline significance (< .06), the epsilon value ($p < .01$) indicates a significant degree of relationship. This rigidity-avoidance relationship was further substantiated by observing that failure to alternate solution on Trial 6 was related (< .10) to avoidant behavior.

Hypothesis 3. Avoidant behavior is directly related to performance disruption after failure stimulation increased. Since this hypothesis was necessarily tested without any reference to these Ss for whom the relationship would be most appropriate (i.e., the nine Ss who left Task 1 at the earliest sign of failure on Trial 5 and therefore had no post-failure performance measure for comparison), the performance disruption-avoidance relationship seems amply supported.

No relationship was noted between any two of the other three variables (expectancy, alternating, and errors) with p levels ranging from .30 to .50.

C. Postexperimental Interviews

The usual function of interviewing Ss following conditional experimentation is

to obtain a relatively independent index of the extent to which the experimental variable has been induced. Viewed simply, *E* would therefore be primarily interested in *S*'s stating verbally that he felt he had failed. However, such a viewpoint overlooks the defensive nature of verbalization. Particularly in the present study utilizing minimal failure stimulation and focusing on avoidant behavior, we did not expect all *Ss* to verbalize feelings of failure. If one considers the interview protocols from a defensive point of view there is no reason to question the successful induction of failure stimulation. However, it is equally clear that there is wide variation in the verbalizations made by *Ss* when describing their reactions to the situation.

Since the results of the objective measures indicated a configural behavioral pattern revolving around failure-avoidant behavior, we attempted to extend this pattern further by noting the verbalizations made by avoidant and nonavoidant *Ss*. It is true, of course, that the nonavoidant *S*, by virtue of his staying with the task, experienced more intensive failure than the avoidant *S*, but even with this difference in experience it appeared potentially useful to note his interview responses. One of the most fruitful questions was "How did you happen to select Task 2 at that point?" i.e., in effect, "Why did you elect to avoid the situation at that point?" When responses of *Ss* who left the task at Trial 5 or 6 (at the earliest signs of failure) were compared with responses of *Ss* who did not leave the task until Trial 12, pronounced differences were noted between the groups. The verbalizations of the avoidant group were primarily of the following variety: "I had done as well as I could," "Didn't think I could do any

better," or "Getting tired of this one." On the other hand the responses of the nonavoidant *Ss* were primarily of the following nature: "Wasn't doing very well so decided I'd better quit," "About time for me to give up," and "Better stop because I was getting worse."

In view of these responses it seemed reasonable to hypothesize that such variation in interpretations of failure might partially account for the configuration of relationships which emerged. That is, the nonavoidant *S* who persisted in looking for other solutions with a minimum of disruption in performance may have behaved in this fashion at least partially because *he interpreted the situation as one in which he was failing*. Conversely, the avoidant *S* might be described as defending himself against failure by behaving in such a way as *to prevent the occurrence of any situation which he might have to interpret as failure*. Further exploration of the possibility of such a relationship will be described in the next two sections.

III. SITUATIONAL INTERPRETATION EXPERIMENT

The situational interpretation experiment (SIE) described here represents a method for measuring how an individual interprets (or perceives) a situation. The primary aim was to measure an individual's tendency to make avoidant interpretations to negative conditions (criticism or failure); the secondary aim was to measure an individual's interpretation of several other situational conditions, e.g., praise and success. Postexperimental data in the first study as well as general agreement in personality theories indicate that a person's perception or interpretation of a situation is of major importance in determining which subclasses

of (the otherwise almost infinite number of) behaviors he will employ in that situation. Since the intervening process of interpretation occupies such a pivotal position in determining responses, it seemed important to attempt to measure such interpretations directly.

It is presently assumed that individuals categorize or summarize their past experiences into systems or concepts or relative constancies (5, 12, 19). Since few situations are ever exactly repeated the individual must generalize in order to adjust effectively to changing conditions or to utilize past experiences effectively. This process of generalization is assumed to result in a tendency to behave in a more similar manner in two or more situations.

Central to the present position is the assumption that, in order to adjust effectively to a situation of failure, an individual must admit that he is doing poorly, that he is in some way inadequate, or that he is, in fact, failing. We assumed that when an *S* interprets a failure situation by thinking "This means I'm not very good at this," that such an interpretation implies an admission of some personal *inadequacy* or *self-negation*. It should be emphasized that we are *not* using the term "inadequacy" in its usual sense which implies behavioral ineffectiveness. In contrast, we mean by "inadequate" that the individual is willing to consider possible weaknesses to admit that he *may* be wrong, thus opening the possibility for modifying his behavior. In considering other possibilities of interpreting a failure situation it is obvious, though nonetheless important, that individuals may avoid making negative interpretations by a host of defensive interpretations, e.g.; *S* may think "This is too difficult for anyone

to do. . . . I'm doing all right however," or "I was just unlucky." Indeed, the entire gamut of defenses emanating from avoidance processes may be considered as functioning to prevent interpretations of personal inadequacy. We will refer to this tendency to avoid making self-negative interpretations to negative conditions as *avoidant interpretation*.

The method outlined here is referred to as a situational interpretation experiment because the individual is confronted with a number of hypothetical situations and then forced to make a choice between two possible interpretations of each situation. His choice of interpretation then becomes the operation for describing the concept the individual used for integrating that particular situational event. *When it has been established that an event, e.g., failure, falls within a given system of integration, e.g., self-evaluation, it is said to be interpreted in a given way.* This method allows a number of observations of each event, in addition to sampling a range of events. It also permits providing various potential interpretations to be used, although in the present form of the SIE we have dealt only with the conceptual system of self-affirmation and negation. (In considering interpretations of positive situations we would not, of course, make the same assumption of avoidance-adequacy equivalence that we made in relation to negative situations, since there is no necessary defensiveness implied in a person's feeling adequate after success.)

A. Method

1. Instructions

All *Ss* were requested to read the face sheet containing instructions which read as follows: "This is a test of how well you can make decisions. The decisions you have to make are not

always easy. In each of the questions you are given a short story about something that is happening. The story gives all the information about what is happening but it does not say anything about the thoughts a person would have if he were in the story. Your task is to complete the story by deciding which thought would follow. In each case one of the two thoughts is a better answer than the other. You are to select the one which is the truest. Read the stories carefully. Ask if you do not understand anything. Make your decisions accurately."

2. Description of Method

The total scale consisted of 30 items,⁴ but the criticism-failure (CF) subscale consisting of ten items below was the primary measure. It will be noted that this subscale is made up of four criticism items, two verbal failure items, and four failure items which may be regarded separately for certain purposes. As the SIE was presented to *S* the same stem, i.e., situation, was sampled a number of times with different alternatives and randomized throughout the test. However, the stem and alternatives are grouped together below with their position in the test indicated by item number. *Self-negative alternatives are indicated by asterisks.*

a. Criticism condition:

You are trying to work out some decisions about a plan of yours. You hope that the plan will be a good one for you. One of the people in charge of the project comes by. He glances at your plans and says that you don't seem to be doing it right, that your plan doesn't seem to be very good. Which is the most correct thought to have?

- Item #7 (a) "He must be mistaken."
(b) "I'm just not up to it."*
- Item #10 (a) "This shows that I'm no good here."*
- (b) "This means that he cannot understand my work."
- Item #15 (a) "He must be right."*
- (b) "This shows he knows nothing about it."
- Item #18 (a) "He knows I'm not doing very well."*
- (b) "It's OK, if there's anything wrong it's only an unlucky mistake."

b. Verbal failure condition:

You are working on a task and hope to do well. Someone looks at your work and says that

you are not doing very well here. Which is the most correct thought to have?

- Item #2 (a) "He must be wrong."
(b) "I'm not doing very well here."*
- Item #29 (a) "I still think it's OK"
(b) "He can see my weaknesses."*

c. Failure condition:

You have been working on a project and you have a strong desire to do well and finish it. At times you have felt that it was giving you a little trouble and now it seems that you will fail at it. Which is the most correct thought to have?

- Item #9 (a) "It is very seldom that I fail."
(b) "This shows that I'm not up to it."*
- Item #13 (a) "This shows that the project must be a very hard one."
(b) "This shows that I might not be so good."*
- Item #19 (a) "This is too difficult for anyone to do."
(b) "If I cannot do this it is bad."*
- Item #27 (a) "I must have been unlucky here."
(b) "This means I'm not very good at this."*

d. The remaining twenty items were divided as follows:

- Praise: four items
Success: four items
Help: four items
Remote praise:⁵ four items
Verbal praise: two items
Peer praise: two items

3. Scoring

It was arbitrarily decided to score the number of "inadequate" or nonavoidant responses in order to arrive at a CF subscale score. The hypothetical range of scores is from 0 to 10, and a low score on this subscale will be referred to as avoidant interpretation. In the case of the other subscales the "inadequate" interpretations were also tallied for scoring.

The present scale is the fourth modification. Apart from modifying items which would yield more equal splits, the major change in this last modification was to substitute an adult as the agent of criticism (and also praise) in place of the peer figure which had been used in earlier

⁴ Copies of the full scale in its complete form are available on request from the authors.

⁵ From S. S. Tomkins, *A model of the human being*. Unpublished. Remote praise refers to a situation containing the subject and at least one other person when the other person is being praised.

forms. The effect of this change is discussed below.

4. Procedure

The SIE was administered to 500 white boys between the ages of 13 and 18 during the summer of 1955. The Ss were drawn from YMCA camps, Boy Scout camps, public and private schools. Approximately half of the Ss were secured on a volunteer basis and half of the Ss were tested in total groups. The testing was carried out in groups ranging in size from 5 to 80. No time limit was set but the time taken ranged from 15 to 30 minutes. In every case the SIE was administered immediately following the Sentence-Completion Method (11) and was followed by a short verbal intelligence test consisting of 20 vocabulary items (27).

B. Results

The distribution of scores on the CF subscale is presented in Table 3 along with the distribution of scores on the praise-success (PS) subscale for comparison. As would be anticipated on an "affirmation-negation" continuum, the scores on the CF subscale cluster at the high (negation) end of the scale while scores on the PS subscale tend toward the low (affirmation) end of the scale. That the CF subscale is internally homogeneous is indicated by the correlation of .34 between the four-item criticism subscale and the failure subscale. The CF subscale also appears to be relatively independent of the PS subscale since the correlation between these two subscales is low ($r = .08$).

For purposes of future investigation it may be noted that the CF subscale correlates .16 with the remote praise and .02 with help (for 500 cases, any correlation of .09 is statistically significant at the .05 level and a correlation of .12 is statistically significant at the .01 level.)

An interesting, but tentative, finding during pretesting an earlier form of the SIE provides some information on possible sources of variance in the criticism

TABLE 3
FREQUENCIES, MEANS, AND STANDARD DEVIATIONS FOR CRITICISM-FAILURE AND PRAISE-SUCCESS SUBSCALES

Score	Criticism-Failure	Praise-Success
0	2	6
1	3	27
2	15	70
3	26	89
4	43	101
5	68	73
6	76	67
7	95	32
8	83	19
9	58	12
10	31	4
Mean	6.48	4.26
SD	2.10	2.03

condition. At an earlier stage the criticism stem consisted of a peer agent rather than an adult as in the present form, and this earlier data indicated that peer criticism elicits fewer self-negative choices than the adult-induced criticism. Such a difference was *not* noted in the agent of praise, so it may be tentatively concluded that interpretations made to criticism are to some extent dependent on the person administering the criticism.

Intelligence. The mean and standard deviation of the verbal intelligence scale were 9.62 and 2.58 respectively. This intelligence scale correlated .10 with the CF subscale ($<.02$) but did not relate to any of the other subscales.

IV. RELATIONSHIP BETWEEN FAILURE-AVOIDANT BEHAVIOR AND FAILURE-AVOIDANT INTERPRETATION

The primary aim of this section was to study the relationship between *avoidant interpretation* and *avoidant behavior* in a failure situation. The secondary aim was to attempt a cross validation of pre-

vious behavioral relationships noted (Section II), using a slightly different method and a different population.

The operation for avoidant interpretation is *S*'s score on the CF subscale of the SIE, with a low score being referred to as avoidant. The operation for avoidant behavior is quite similar to that used in Section II—the trial at which *S* avoids the problem situation in which he is experiencing increasing failure by selecting a different goal.

Measures of interpretation and behavior were obtained according to the method to be described in order to test the hypothesis that failure-avoidant interpretation would be associated with failure-avoidant behavior. An attempt was also made to obtain other behavioral measures (used in Section II) and other test measures, e.g., intelligence, in order to clarify further the nature of avoidant behavior. However, the matter of crucial importance is the utilization of avoidance as reflected in interpretation and behavior.

A. Method

1. Subjects

The *Ss* were 66 high school boys between the ages of 13 and 18, attending two summer schools.

2. Materials

a. Task 1 consisted of a multiple-choice type of symbol-substitution test. Each trial required approximately 15 seconds.

b. Task 2 was a card-sorting task consisting of a sorting box identical to that used in II and of three demonstration packs and six task packs of eight cards each. In all cases at least one concept was relevant for sorting. When a concept was relevant, a correct sort always produced four groups of two cards each. The demonstration packs were used to insure that *S* understood the principles of grouping. The first three task packs used on the first three trials contained five concepts (letter, object, num-

ber of dots, form, and color) but only one (letters) was relevant for grouping. For example, in the first pack all eight cards contained the same object (shoe), the same number of dots (three), the same form (triangle), and the same color (green), but differed with respect to letters (two cards with L, two with M, two with N and two with O). On the second pack the four nonrelevant concepts were again constant throughout but different from Pack 1. In other words, on the first three packs there was interpack variation, but no intrapack variation for all concepts except the relevant concept, namely letters. The fourth pack contained one relevant concept (letters), but the other four concepts changed on Trial 5 to indicate to *S* that a change in the situation had occurred. Simultaneously failing scores were given. Packs 5 and 6 contained five relevant concepts which could be used for sorting and were alternated from trial to trial, after Trial 5.

c. Sets of "norms" ostensibly used to compute *S*'s score.

d. Score sheets and record slips used to inform *S* of his score and make his choice of Task 1 or Task 2 after each trial.

3. Procedure

The group tests were administered in the following order: (a) The Sentence-Completion Method (11), (b) the Situational Interpretation Experiment, (c) the Thorndike-Gallup Verbal Intelligence scale (27). During the week following the group tests, each *S* was seen individually. In this experimental procedure *S* was read a set of instructions designed to emphasize the importance of succeeding on Tasks 1 and 2. The problems were presented as measures of decision-making ability. So that *S* would understand the value of any score obtained, he was informed in both Tasks 1 and 2 as follows: "In problems like this you will improve with practice; that is, each time you do it, you should get a better or a higher decision making score . . . your score should get higher as you go along. You should get up to a score of between 30 and 35 with practice which would be a good score for high school boys." The score pattern for Task 1 was 6, 14, 20, 31. After the fourth trial on Task 1, *E* said to *S*, "It looks as though you have learned to do this very well . . . you are doing very well on this problem, so let us take a look at Task 2 now (indicate). Later you might want to come back to this problem if you like." At this point *E* and *S* moved to Task 2, plainly marked at the opposite end of the table. The nature of the sorting task was clearly demonstrated to *S* before beginning. On the first three trials only

one cue was relevant for sorting, but at Trial 5, all five cues were relevant and simultaneously the scores leveled off and began to decrease. The score pattern for Task 2 was 8, 15, 21, 22, 22, 21, 19, 18, 17, 15. After each trial *S* was handed a slip containing his score, a space to check his preference for Task 1 or Task 2 on the next trial, and his expectancy for the next trial. The experiment was continued until *S* selected Task 1 or until the twelfth trial.

Although the present method was intentionally similar in most respects to that of the college study (Section II) there were two differences: (a) The cards were so constructed that *S* could not alternate his solution until the fifth trial, a modification designed to obtain a better measure of postfailure behavioral alternation (in the first study *Ss* varied in their prefailure alternating, thereby making postfailure analysis difficult). In practice, however, only a small proportion of *Ss* alternated after failure, so that further revision of the method is indicated in this respect. (b) Each *S* experienced success on the alternate task (Task 1), so that avoidance in the present study indicated not only avoiding the failure situation imposed in Task 2, but also going back to a previously successful problem or goal (in the first study some *Ss* switched to the alternate goal before Trial 5, apparently because they were curious about the nature of the other task; so that it seemed worth while to give each *S* a similar experience with the alternate task). Even with this modification, however, it was necessary to eliminate 12 *Ss* from analysis because they selected the alternate goal before the fifth trial.

B. Results

1. Avoidant Behavior and Interpretation

Table 4 indicates a significant relationship between avoidant interpretation and avoidant behavior. The interpretation score in Table 4 is based on the criticism-failure subscale score. When a similar table based only on the failure subscale score is considered, the relationship is also significant ($<.01$) even though this failure subscale consists of only four items.

We noted previously that an individual may avoid an inadequate interpretation of a negative event in a variety of ways, and actually the avoidant alternatives on the criticism-failure subscale

TABLE 4
RELATIONSHIP BETWEEN AVOIDANT INTERPRETATION AND AVOIDANT BEHAVIOR

Behavior	Avoidant Interpretation (1-7 on CF subscale)	Nonavoidant Interpretation (8-10 on CF subscale)
Avoidant (4-7)	19	6
Nonavoidant (8-13)	11	18

Corrected $\chi^2 = 6.401$; $p = <.02$.

present different defenses to avoid accepting the negative event. Therefore, we made an item analysis of the ten criticism-failure subscale items (and later on the other 20 items also) to note whether the selection of a given alternative related to either avoidant or nonavoidant behavior. In such an analysis the selection of an avoidant alternative may provide a good predictor which will discriminate between extreme behavioral groups, but selection of the opposite alternative on the same item need not be equally discriminating.

Seventy per cent or more of the *Ss* who selected the following alternatives were behavioral avoidant *Ss*:

27. (Failure) "I must have been unlucky here."
19. (Failure) "This is too difficult for anyone to do."
18. (Criticism) "It's OK. If there is anything wrong, it's only an unlucky mistake."

Seventy per cent or more of the *Ss* who selected the following alternatives were behavioral nonavoidant *Ss*:

10. (Criticism) "This shows I'm no good here."
11. (Failure) "This shows I might not be so good."

In addition it was further noted that the following two items from the help subscale were selected by at least 70 per cent of the behavioral nonavoidant *Ss*:

12. (Help) "This shows how bad he thinks I am."
4. (Help) "It means I probably couldn't do it myself."

These results seem to lend further, more specific, support to our initial hypotheses regarding the processes underlying avoidant behavior. The alternatives selected by the nonavoidant *S* imply a willingness to lower his self-evaluation in reaction to failure or criticism while the alternatives selected by the avoidant *S* imply defenses against any such lowering of self-evaluation.

2. Other Relationships

No relationship was noted between verbal intelligence scores and avoidant behavior. There was not sufficient variation on behavioral alternation to replicate the rigidity-avoidance finding. However, the relationship between avoidance and expectancy was again noted ($<.05$) which provides another cross validation for this finding on a different population.

V. RELATIONSHIP BETWEEN AVOIDANT INTERPRETATION AND POSTCRITICISM SELF-EVALUATION

The primary aim of this section was to investigate the relationship between avoidant interpretation and self-evaluation following criticism. A secondary phase, identical in method to the primary phase, concerned interpretation and self-evaluation under praise conditions. In both their verbal comments (II) and interpretation alternatives selected (IV), behavior avoidant *Ss* apparently utilized defenses against lowering their self-evaluation following the occurrence of a negative condition. Therefore, in order to test this relationship more precisely, we employed a direct measure of

self-evaluation as a dependent variable in the present study. Contrasting groups, representing avoidant and nonavoidant *Ss* on the CF subscale of the SIE, were individually placed in a performance situation and *Ss* self-evaluation response following *E's* criticism on a given trial was measured. It may be noted that the hypothesis tested represents an explicit effort to validate the criticism-failure subscale.

It was hypothesized that avoidant *Ss* would be more likely to hold constant or increase their postcriticism evaluations while nonavoidant *Ss* would be more likely to decrease their postcriticism evaluations.

We have stated earlier that we would not make the assumption of adequacy-avoidance equivalence with respect to the praise or success conditions. Indeed, if one maintained a relative definition of avoidant interpretation, i.e., *S's* implied rejection of the event, we might assume that an inadequate interpretation of praise would be avoidant. In light of this consideration, we applied the same approach outlined above to the problem of "praise-avoidant" interpretations and postpraise self-evaluation for preliminary exploratory purposes.

A. Method

1. Subjects

A total of 61 *Ss* were selected as follows:

Criticism condition. Two groups of *Ss* were selected on the basis of their criticism-failure subscale score. Eight *Ss* scoring between 0 and 2 were considered the extreme avoidant group and 18 *Ss* scoring between 8 and 10 were considered the nonavoidant group.

Praise condition. Two groups of *Ss* were selected on the basis of their praise-success subscale score. Seventeen *Ss* scoring between 8 and 10 were considered the avoidant group and 18 *Ss* scoring between 0 and 2 were considered to be the nonavoidant group.

2. Praise Condition

Analysis of the praise condition indicated that "praise avoidant" interpretations were significantly ($<.05$) related to failure to increase postpraise evaluation. No relationship was found between expectancy and evaluation in either of the subscales to absolute evaluation, a measure which will be considered in the next section.

VI. RELATIONSHIP OF AVOIDANT INTERPRETATION TO OVEREVALUATION OF PERFORMANCE AND ANXIETY

The study described in the previous section was designed to minimize all feedback to *S* except *E*'s comment, i.e., *S* received no "knowledge of results" or score on trials preceding *E*'s comment. The present section was initially aimed to study the role of avoidant interpretation as a determinant of *S*'s reaction to criticism administered in relation to a given score pattern. More specifically, contrasting modes of information were presented to *S* by *E*'s interjecting criticism (negative evaluation) in relation to an increasing score pattern (positive accomplishment). We reasoned that *S*'s implicit choice between such contrasting modes as indicated by his subsequent self-evaluation should reflect interpretation tendencies, i.e., avoidant *S*s' "accepting" the positive score rather than the negative evaluation, and conversely for nonavoidant *S*s.

However, difficulties were encountered in developing a score pattern which both (a) induced feelings of accomplishment and (b) appeared sufficiently ambiguous that the negative evaluation given in relation to these scores was not immediately rejected. The score pattern finally adopted was more ambiguous than posi-

tive, i.e., *S* experienced slight improvement.

An inspection of data using this score pattern indicated an unanticipated finding: namely, that the effect of the interpretation variable was of major importance on the initial self-evaluation following the first (rather low) score (see Table 6). This interpretation-produced difference was in evidence throughout the six trials and therefore masked any possible effects of *E*'s evaluative comment. Therefore, we will present only the method and results based on the first three (pre-evaluation) trials since the experimental induction had a negligible effect.

This study will describe (a) the relationship of avoidant interpretation to both absolute evaluation and tendency to overevaluate performance, and (b) relationship between avoidant interpretation and anxiety as measured by the Taylor anxiety scale (24).

As the study was carried out, *S*s were assigned to two conditions since praise was also employed in addition to criticism. However, since neither evaluation was effective and the pre-evaluative score patterns (for criticism—11, 12, 14 and for praise—13, 15, 14) did not produce group differences in either self-evaluation or performance during the first three trials, these conditions were combined for purposes of present analysis.

A. Method

1. Subjects

The *S*s were 80 unselected high school boys who were attending a parochial summer school for purposes of raising their grades, rather than because of failing course work.

2. Materials

Eight symbol-substitution problems from the Rotter-Jensen group level-of-aspiration test (20). The lower line of symbols was removed so that

each problem contained 25 symbols to be completed. Evaluation and expectancy after each trial were recorded using the same materials described in Section V.

3. Procedure

The Ss were divided into two groups of 42 and 38 for purposes of varying group testing order. In the first group the testing order was: Sentence Completion Method (11); Taylor Anxiety Scale (24); Situational Interpretation Experiment; and a short verbal intelligence test (27). In the second group the order was identical except that the order of SIE and anxiety scale was reversed. Following the group testing, Ss were assigned to one of four subgroups to counterbalance for two conditions and two Es by matching on interpretation variable.

Experimental procedure was similar to that described in Section V except that S was told he would have a given amount of time in which to do each trial. This procedure permitted E to stop each S at the same predetermined number of symbols for a given trial. Thus, S became aware of his score without the necessity of E's announcing it, since the symbols were plainly numbered in consecutive order. The maximum hypothetical score was 25. The predetermined score patterns for the first three trials in both conditions were rather low (11, 12, 14, and 13, 15, 14) so that in effect both patterns may be considered as representing low, possibly failing, scores.

4. Measures Employed in Analysis

a. Personality measures.

1. CF subscale score
2. Anxiety scale score

b. Behavioral measures.

1. Absolute evaluation: actual box checked following a given trial (possible range from 0 to 12)
2. Corrected rate: number of correct symbols per minute. This was used to compute—
3. Over- and underevaluation: In order to obtain a measure of evaluation relative to performance which would permit identification of Ss who overevaluated or underevaluated their performance, the following procedure was employed: median cutting points were established for both the corrected rate and total evaluation scores for the first three trials. Using these two measures, Ss were appropriately placed into one of the four resulting quadrants. Therefore, Ss in the quadrant of high evaluation and low rate of per-

formance were referred to as "overevaluators," and Ss in the quadrant of low evaluation and high rate of performance were referred to as "underevaluators" (Ss in both conditions were grouped together, since these measures are based on trials preceding E's evaluative comment).

B. Results

1. Relationships Between Personality Measures

Pearsonian correlation coefficients computed between anxiety and criticism-failure subscale score yielded a correlation of $-.29$, which is statistically significant at less than the $.01$ level. That is, high anxiety is associated with failure-avoidant interpretation. This relationship must be qualified since a testing-order effect was noted. When SIE was given first, $r = -.55$ ($<.01$ for $N = 38$), and when SIE was given last, $r = -.20$ (not significant for $N = 42$).⁶

2. Relationship Between Avoidant Interpretation and Evaluation

a. *Absolute evaluation.* In order to note the effect of interpretation upon absolute evaluation, the initial evaluation scores of 22 avoidant Ss in the lower quartile on the CF subscale were compared with the scores of 21 nonavoidant Ss in the upper quartile on the same subscale.

As Table 6 indicates, there is a significant tendency for the avoidant Ss to give higher self-evaluation following the first score (a relationship which held throughout all six trials). That this relationship is not limited only to the extremes is indicated by the correlation between evaluation score on first trial and criticism-failure subscale score which was $-.21$ $r < .05$ with $N = 80$).

⁶ Recently we have had an opportunity to replicate this finding, but in two samples the correlation between CF score and anxiety failed to reach significance.

TABLE 6
RELATIONSHIP BETWEEN AVOIDANT INTER-
PRETATION AND FIRST-TRIAL EVALUATION

Evaluation	Avoidant Interpretation (1-4 on CF subscale)	Nonavoidant Interpretation (8-10 on CF subscale)
Low	6	16
High	16	5

Corrected $\chi^2 = 8.383$; $p = < .01$.

b. Over- and underevaluation. Since the tendency to give a high evaluation carries more personality-related significance if it occurs in relation to low performance, the role of avoidant interpretation was inspected as it related to over- and underevaluation. Following the procedure previously described for determining these two extreme groups of evaluators left only 29 Ss in both groups, which required a median rather than a quartile split on the interpretational variable.

As Table 7 indicates, the CF score is highly related to tendency to under- or overevaluate one's performance. This tendency for the avoidant S to overevaluate his performance is consonant with the general characteristics of this group in other respects. There was no relationship whatsoever between either the PS subscale or the anxiety scale to over- and underevaluation.

TABLE 7
RELATIONSHIP BETWEEN AVOIDANT INTER-
PRETATION AND OVER- AND UNDER-
EVALUATION

Group	Avoidant Interpretation (1-7 on CF subscale)	Nonavoidant Interpretation (8-10 on CF subscale)
Underevaluators	2	12
Overevaluators	12	3

Corrected $\chi^2 = 10.029$; $p = < .002$.

Nonavoidant Ss tended to shift their evaluations more ($< .10$) than did avoidant Ss. No relationship was noted between avoidant interpretation and expectancy.

VII. DISCUSSION

A. Avoidant Behavior

In considering the results described in the preceding sections it appears that the behavior-avoidant S employs defenses which at once (a) distort his evaluation of his capability (setting excessively high goals and overevaluating his performance), and (b) restrict the reception of potentially useful information from the environmental situation (maintaining his post-criticism self-evaluation and making fewer shifts in evaluation). This defensive pattern may partially account for his relative inability to try out other solutions, as well as for his disrupted performance—this latter consequence resulting from the fact that he does receive *some* feedback; and when such negative information is received, it is poorly assimilated. Such disruption in transforming evaluational feedback into more adaptive behavior is very similar to the behavior of high-anxious Ss in their reactions to failure as described by Mandler and Sarason (15). Restriction in the assimilation of information has also been described as characteristic of a group of "overcontrollers" described by Block and Thomas (3) who regard such a pattern as "metastable."

In light of this general relationship between personality factors and problem-solving behavior found here and in other investigations, it would seem that subsequent research might be appropriately directed toward specifying these defensive patterns more specifically. Once such patterns have thus been objectively delineated, work may then proceed on the crucial problem of determining the conditions which give rise to the development of such avoidant patterns.

B. Avoidant Interpretation

The results obtained using the interpretation measure indicate that an individual's interpretation or perception of a negative situation is an important determinant of subsequent responses. Previous attempts to measure the nature of an individual's interpretation of a situation have been primarily carried out using indirect approaches, e.g., TAT. That is, the analysis of a TAT protocol, whether for clinical or experimental purposes, frequently consists of inferences regarding how a person perceives a

certain situation. Whether one is dealing with maladjusted patients or subjects in a conditional experiment, it is equally true that the responses of such persons are not comparable unless the event has been interpreted in a similar way. An investigation described by Atkinson (1) which aimed at determining the picture properties eliciting certain interpretations is a case in point.

In light of these considerations we believe that it is necessary to distinguish between the measurement of (a) an individual's tendencies to interpret or organize a situation, e.g., a TAT picture, and (b) his available reactions to this event once it has been interpreted in a given way. That is, if one wanted to measure an individual's prepotent organizational tendencies, he should probably select a highly ambiguous stimulus; however, if the investigator desired to assess the individual's reactions to certain events, it might be more appropriate to select more highly structured stimulus materials. This latter procedure may be exemplified by a hypothetical procedure of presenting Card 6BM in the Murray series (the "mother-son" card) with the structuring sentence, "This man has just lost his job and is talking to his mother about it." It should be made clear that we do not assume any necessary one-to-one relationship between interpretation and behavior, but rather would maintain that once the interpretation of the situation has been measured (and, if possible, controlled) that the behavior may be studied more accurately.

The present use of a forced-choice single dimensional method for assessing interpretations is limited in that the reference system employed (adequacy-inadequacy in the present case) may not be equally appropriate for all Ss. This difficulty is evidenced by the relative lack of predictive precision whenever it was necessary to split Ss at the median on the interpretation measure (see Table 4). However, in cases where the number of Ss permitted the use of extreme avoidant groups (e.g., 0-2 and 8-10 as in Table 5) the resulting efficiency was greatly increased. In other words, it seems likely that for those Ss who score in the 3-7 range, the reference system being employed is not especially relevant. However, the methodological paradigm of the SIE should lend itself to studying the probability that other conditions, e.g., authority situations, will be interpreted by other reference systems, e.g., aggression. Tentative evidence for the utility of the SIE measure for other conditions is noted in the positive results reported with the praise-success subscale (Section V).

We have also briefly noted the effect of the person or the agent in the situational stem. That is, the differences in interpretation pro-

duced by peer-induced and adult-induced criticism suggest that the method might be useful for measuring indirectly the meaning or significance which a given figure in the situation may have for one individual or for a group.

The implications of Sections V and VI to research in the area of individual variation in reaction to criticism, failure, praise, and success may be briefly considered. Experimenters in the area of social motivation have for some time been concerned with the fact, noted by Grace, that "that which might act as blame for one individual might act as praise for another" (9, p. 77). Utilizing constructs such as introversion-extroversion (26) and ascendancy-submission (25) previous attempts have focused on accounting for performance variation, implicitly assuming that some change in self-evaluation has intervened. In using the present approach anchored in this intervening process, we have indirect evidence (Sections II, V) to support the contention that Ss who refuse to lower their evaluation following failure or criticism perform more poorly.

That interpretation scores on the criticism-failure subscale were associated with initial absolute evaluation (Section VI) is not surprising when it is noted that this evaluation followed an initial, fairly low score which may have been interpreted as failing. These initial differences in absolute evaluation masked any hypothesized changes in relative evaluation—which leads to the speculation that the interpretation variable appears to have its maximal effect in reaction to the *first cue* provided, in this case, a low score. A similar primacy of effect has been noted in connection with differences in response attributable to an anxiety variable (16), in that differences between anxiety groups were "diluted" after the first trial.

In considering the findings from Sections V and VI it is important to recall that in the former, S received no scores or other evaluational information until E interjected a comment, and prior to this, there were no differences between subgroups in absolute evaluation. In Section VI when scores were given, differences in evaluation were present from the outset. In attempting to synthesize findings from these two studies, it would appear that the nonavoidant Ss are sensitized to the *initial evaluational cues* in a new situation whether such cues come from their scores or from E. This heightened sensitivity characteristic of nonavoidant Ss may be viewed in contrast to the previously described defensively-restrictive characteristic of the avoidant Ss. In overevaluating their performance (Section VI) the avoidant Ss bear a certain resemblance to the overly confident Ss described by Block and Petersen (2) as rigid and dogmatic. Conversely a certain similarity in person-

ality pattern is noted between our nonavoidant Ss and the overly cautious Ss described by these authors as tending toward introspection and self-abasement. The present referents for avoidant and nonavoidant interpretation might also be considered as reflecting different means of dealing with blame-assignment (17).

C. Relationship of Avoidant Interpretation to Intelligence

The relationship between low intelligence and avoidant interpretation must be qualified. But, even if such a low-order relationship is verified, this does not necessarily detract from the value of the interpretation measure. That is, it may well be true that one of the possible antecedents for the adoption of avoidant interpretation is a deficiency in problem-solving ability. That intelligence is not the only factor involved in avoidant interpretation is to be noted in the fact that it is not related to behavioral avoidance (Section IV).

D. Concluding Comments

From a conceptual point of view our findings represent what Cronbach and Meehl have called a "nomological net" (7) which is anchored primarily in the behavior of failure-avoidance and secondarily in the behavioral measure of self-evaluation. The present results are implicit evidence for the value of considering the psychological situation in attempting to predict the reactions occurring in that situation.

Certain shortcomings and limitations of the present approach should be mentioned. In our present exploratory attempt to measure interpretations we have used a relatively imprecise instrument, with the disadvantage that the avoidant alternatives are not necessarily homogeneous regarding defenses. We have not demonstrated any independent criteria for assessing the utility of failure-avoidant behavior although the postexperimental interviews indicated that the failure stimulation was successfully induced. Further, we have implicitly assumed linearity in all of our measures in our use of chi-square analysis. However, in this exploratory effort we hoped that the disadvantages in lack of precision would be outweighed by the advantages of gaining a better understanding of the problem.

The reader has probably noted that in our attempt to demonstrate interpretation-produced variation in behavior we have explicitly omitted any consideration of certain other personality variables of obvious importance, e.g., psychological need or motive. The relationship of interpretational variables to other constructs such as psychological need and generalized ex-

pectancy (18) and their combined effect are problems deserving further concerted investigation.

To illustrate the conceptual complexity of this problem, Atkinson's recent description of an attempt to measure fear of failure by scoring imaginative productions may be noted (1). Although this fear-of-failure measure appears to be operationally similar to the present avoidant interpretation measure, Atkinson conceptualizes this construct as a "negative" motive with properties opposite to those of achievement motive. We cite this example not as an "opposing theory," since we do not feel justified in presenting a complete theoretical conception of failure-avoidance at this time, but rather to point out the necessity for further investigation. We feel strongly, however, that such investigation requires relevant behavioral anchor points in order to develop theoretical conceptions for the understanding of avoidant processes. In the present investigations the measures of avoidant behavior and self-evaluation appear to have served this purpose.

VIII. SUMMARY

A series of hypotheses derived from the problem-solving behavior and verbalizations of individuals employing avoidant mechanisms were tested and generally supported. These findings, which appear to be consonant with general observations regarding the operation of defensive processes, may be summarized in terms of two constellations of relationships.

The Ss who utilized failure-avoidant behavior in a problem-solving situation were found to: (a) set higher goals, (b) use fewer alternative solutions in attempting to solve the problem, and (c) perform less effectively after failure. These relationships were noted in two identical studies using male college Ss. Following leads which emerged from the postexperimental interviews, a measure of an individual's tendency to avoid interpreting negative situations in terms of personal inadequacy was constructed.

In employing this interpretational measure with high school boys, Ss who

made avoidant interpretations of failure and criticism were found to (a) avoid failure in a problem-solving situation, (b) maintain their self-evaluation after criticism, (c) state higher evaluations following a low "failing" score, and (d) over-evaluate their performance.

Though these generalizations must be limited to college and high school male populations studied, the results are based

on separate studies using similar measures, and in at least one case—behavioral avoidance and high goal-setting—we were able to replicate this relationship in both populations. These relationships appear to integrate a number of experimental and clinical observations between personality characteristics and problem-solving behavior under negative conditions.

REFERENCES

1. ATKINSON, J. W. Explorations using imaginative thought to assess the strength of human motives. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Univer. of Nebraska Press, 1954.
2. BLOCK, J., & PETERSEN, P. Some personality correlates of confidence, caution, and speed in a decision situation. *J. abnorm. soc. Psychol.*, 1955, **51**, 34-41.
3. BLOCK, J., & THOMAS, H. Is satisfaction with self a measure of adjustment? *J. abnorm. soc. Psychol.*, 1955, **51**, 254-259.
4. CAMERON, N., & MAGARET, ANN. *Behavior pathology*. New York: Houghton Mifflin, 1951.
5. CANTRIL, H. Ethical relativity from a transactional point of view. *J. Philol.*, 1955, **52**, 677-687.
6. CHILD, I. L., & WATERHOUSE, I. K. Frustration and the quality of performance: III. An experimental study. *J. Pers.*, 1953, **21**, 298-311.
7. CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281-302.
8. ERIKSEN, C. W., LAZARUS, R. S., & STRANGE, J. R. Psychological stress and its personality correlates. II. The Rorschach test and other personality measures. *J. Pers.*, 1952, **20**, 277-286.
9. GRACE, GLORIA L. The relation of personality characteristics and response to verbal approval in a learning task. *Genet. Psychol. Monogr.*, 1948, **37**, 73-103.
10. HOVLAND, C. L., & KENDLER, H. H. The New York University conference on human problem solving. *Amer. Psychologist*, 1955, **10**, 64-68.
11. HUNT, D. E., & SCHRODER, H. M. An objective system for verbal behavior analysis and its application. *J. abnorm. soc. Psychol.*, in press.
12. KELLY, G. A. *The psychology of personal constructs*. Vol. 1. *A theory of personality*. New York: Norton, 1955.
13. LAZARUS, R. S., & ERIKSEN, C. W. Effects of failure stress upon skilled performance. *J. exp. Psychol.*, 1952, **43**, 100-105.
14. MCCLELLAND, D. C., & APICELLA, F. S. A functional classification of verbal reactions to experimentally induced failure. *J. abnorm. soc. Psychol.*, 1945, **40**, 376-390.
15. MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 166-173.
16. MANDLER, G., & SARASON, S. B. The effect of prior experience and subjective failure on the evocation of test anxiety. *J. Pers.*, 1953, **21**, 336-341.
17. ROSENZWEIG, S. An outline of frustration theory. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. Vol. 1. New York: Ronald, 1944. Pp. 379-388.
18. ROTTER, J. B. *Social learning and clinical psychology*. New York: Prentice-Hall, 1954.
19. ROTTER, J. B. The role of the psychological situation in determining the direction of human behavior. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Univer. of Nebraska Press, 1955.
20. ROTTER, J. B., FITZGERALD, B. J., & JOYCE, J. N. A comparison of some objective measures of expectancy. *J. abnorm. soc. Psychol.*, 1954, **49**, 111-114.
21. SEARS, PAULINE S. Levels of aspiration in academically successful and unsuccessful children. *J. abnorm. soc. Psychol.*, 1940, **35**, 498-536.
22. SEARS, R. R. Initiation of the repression sequence by experienced failure. *J. exp. Psychol.*, 1937, **20**, 570-580.
23. SCHRODER, H. M., & ROTTER, J. B. Rigidity as learned behavior. *J. exp. Psychol.*, 1953, **44**, 141-150.
24. TAYLOR, JANET A. A personality scale of

Psychological Monographs: General and Applied

Pure-Tone Thresholds Following Stimulation by Narrow-Band Filtered Noise¹

JOHN L. FLETCHER

1st Lt., MSC, USA

Army Medical Research Laboratory, Fort Knox, Kentucky²

THE literature regarding auditory fatigue is abundant. Without attempting an exhaustive review of the literature, the studies of Caussé and Chavasse (3), Davis *et al.* (5), Epstein (7), Harris (10, 11, 12), Hirsh and Ward (14), Rawdon-Smith (22), Rawnsley and Harris (23, 24), and Rosenblith (25), Rosenblith, Galambos, and Hirsh (26), and Rosenzweig and Rosenblith (27) show that much effort has been exerted to relate such factors as the frequency, intensity, and duration of exposure of test sounds to poststimulatory auditory thresholds.

Evidence has appeared indicating that auditory poststimulatory effects can be bidirectional, i.e., either an increase or a decrease in auditory sensitivity can follow intense auditory stimulation. Several investigators have reported what has come to be called "sensitization" (1, 2, 4, 5, 8, 12, 13, 14, 15, 16, 21) following stimulation. It is here defined as a lowering of the normal auditory threshold (or conversely, an increase in normal auditory acuity). Some investigators (1, 13, 16) report that sensitization varies with the frequency of the test tone or with the frequency of the exposure tone. Bronstein (1) found that sensitiza-

tion was less for low frequency tones and increasingly greater for higher frequencies. Hughes (15) noted maximum sensitization at 300-500 cps. He also reported that one of his three extensively trained subjects showed sensitization to a pure tone after a white-noise stimulus.

One theoretical formulation has been offered to explain sensitization. Hughes suggested *post-tetanic-potential*³ as a possible answer. He believes that the continued stimulation of the auditory system by the exposure sound tetanizes the relevant nerve fibers, provided the sound is of sufficient intensity. Eccles (6) found that continued tetanization of nerve fibers increased the poststimulatory nerve potential. Hughes concluded that sensitization was due to an increase in poststimulatory potentials from the stimulated auditory nerve fibers.

If post-tetanic-potential is a valid explanation for sensitization, then a broad-band noise source used as an exposure sound at an appropriate intensity should result in sensitization for tones within the frequency limits of the noise-band stimulus. On the other hand, if the noise were fed through the proper acoustic filters and a narrow gap removed from it, there should be less or perhaps no sensitization for frequencies within the filtered gap because less tetanization should result in a decreased sensitization.

¹ This monograph is a portion of a doctoral dissertation done at the University of Kentucky under the direction of Dr. James S. Calvin. The author wishes to acknowledge the invaluable contributions of Dr. Calvin and the rest of the doctoral committee and of Dr. Max S. Schoeffler who assisted in the development of the theoretical analysis of the results of this study.

² The opinions or conclusions contained in the present report are those of the author. They are not to be construed as reflecting the views or endorsement of the Department of the Army.

³ Eccles (6) defines post-tetanic-potential (PTP) as the "... relatively prolonged increase in response that occurs after a junctional region has been subjected to repetitive orthodromic activation. ..." In other words, PTP is an observed, persistent increase in neural activity at a synapse following tetanic stimulation.

If a filtered noise is used as an exposure sound and pre- and postexposure thresholds are determined for tones within the filtered gap, reasoning suggests that the receptors associated with the frequency area within the filtered gap would receive less stimulation during exposure and, therefore, be characterized by a different poststimulatory response than for those of frequencies outside the filtered gap. An unfiltered noise stimulus could serve as a control for testing the above hypothesis.

The purpose of the present study was to determine the effect of filtered and unfiltered noise stimuli at two sensation levels upon the poststimulatory thresholds for pure tones located in and outside the filtered gap.

I. APPARATUS

The noise stimuli were recorded on high-fidelity magnetic tape using an Ampex 600 series recorder. The recording was made at a tape speed of 15 inches per second. The white-noise source was a Gerbrands Noise Generator, P.A.L.,

type 422. A gap extending from 100 to 600 cps was filtered from the noise by passing the signal through two series-connected Krohn-Hite 350-A ultra-low-frequency rejection filters which were placed ahead of the recorder. The filters were calibrated by a series of pure-tone input signals from a Hewlett Packard model 200-C audio-oscillator.

A block diagram of the apparatus used is shown in Fig. 1.

II. SUBJECTS AND PROCEDURE

Seven subjects (Ss) were used. All of the Ss but one had had prior experience in attending to auditory stimuli in experimental studies. All Ss had normal monaural acuity (best ear). Normal is here defined as acuity of not more than 10 db below normal at the octave frequencies 125-8000 cps. An otological examination was given all prospective Ss and none was accepted who showed any evidence of ear, nose, or throat pathology. An alternate loudness balance test was used to screen for recruitment. None of the Ss used in this study showed recruitment.

1. KROHN-HITE MODEL 440-A OSCILLATOR
2. AMPEX MODEL 400 TAPE RECORDER
3. R.W. CRAMER INTERVAL TIMER
4. FISHER MODEL 50C PRE-AMPLIFIER
5. FISHER MODEL 50A AMPLIFIER
6. CLICKLESS SWITCH
7. HEWLETT-PACKARD MODEL 350 ATTENUATOR
8. WESTON MODEL 769 VTVM
9. DUMONT TYPE 218 CRO
10. UTC LS34 MATCHING TRANSFORMER
11. PERMOFLUX PDR-10 EARPHONES WITH DOUGHNUT CUSHIONS
12. LANGE OF LEXINGTON, WAXED PAPER RECORDER AND HAND HELD SWITCH

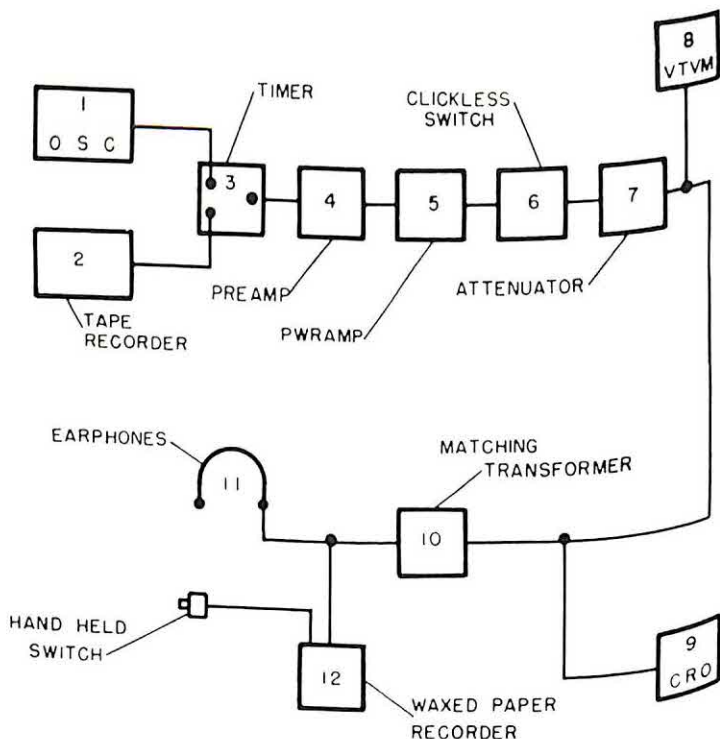


FIG. 1. Block diagram of apparatus.

All testing was done in a sound-treated room with an average ambient noise level of 45 db. Ss wore earphones with doughnut cushions which provided about 20 db attenuation of environmental sound. Thus, the Ss' thresholds were determined against a 25 db background (45 db ambient noise minus 20 db earphone attenuation).

The serial method of limits was used to determine all thresholds. Each threshold was fixed by 4 crossings, 2 ascending and 2 descending series.

The Ss were stimulated monaurally and monaural thresholds were taken.

Familiarization training was given to each S before the experiment was begun. This training consisted of 3 complete audiograms, 5 trials to determine thresholds for the study, and 2 white-noise threshold determinations.

As soon as the S came into the testing room the earphones were placed upon his head and he was seated. His threshold for the test-tone he was to be exposed to for the experimental session was then determined and he was exposed for 4 minutes to the filtered or the unfiltered noise. At the end of the 4-minute exposure period the noise was terminated, the test-tone circuit was switched on, and the experimenter began determining the S's postexposure thresholds for the test-tone. The order of presentation of the variables of intensity, test-tone frequency, and filtered or unfiltered noise exposure was randomized.

III. EXPERIMENTAL VARIABLES

Two sensation levels were used, one of 100 db SL, the other of 20 db SL.

The Ss were exposed to 2 different exposure sounds, a filtered and an unfiltered noise. The filtered noise had a gap filtered from 100 to 600 cps with a center frequency maximum attenuation of 25 db. The unfiltered noise stimulus was broad-band noise with a fairly flat spectrum. Duration of exposure to the noise stimuli was 4 minutes, with 4 hours intervening between experimental sessions.

The test-tone frequencies were as follows: one tone of 255 cps as the center frequency (which corresponded to the center frequency of the filtered gap), a tone of 350 cps which was above the center frequency, and one of 162 cps which fell below the center frequency.

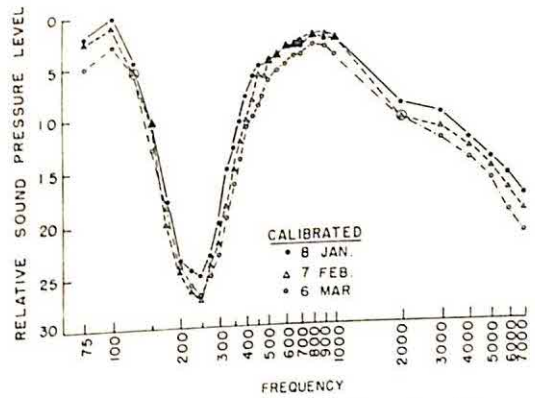


FIG. 2. Calibration of the filtered noise.

The 162- and 350-cps tones in the filtered gap were attenuated approximately half as much as the 255-cps tone. Then signals of 600-cps and 100 cps were selected. These signals were chosen because they fell at the lower and upper edges of the filtered gap. Test-tone frequencies were selected by examining the calibration curve of the filtered noise (see Fig. 2).

Five postexposure thresholds were taken within time intervals of 45 seconds, 90 seconds, 2 minutes, 4 minutes and 8 minutes after termination of the exposure sound.

IV. RESULTS

The major finding of this study is that Ss stimulated with the unfiltered noise consistently had higher postexposure thresholds for pure tones within the filtered gap than did those who were stimulated with the filtered noise. This suggests that the unfiltered noise produced auditory fatigue while the filtered noise resulted in sensitization. The above possibility was tested by the paired replicates test (23). Results of the statistical analysis of the data may be seen in Table 1. Figs. 3, 4, and 5 present the graphical analysis of the data.

When unfiltered noise at 100 db SL was presented, poststimulatory sensitivity to all test-tones was lowered (Fig. 4).

TABLE 1
STATISTICAL ANALYSIS OF DATA

A. $2 \times 2 \times 5 \times 7$ Over-all Analysis of Variance

An analysis of the effects of 2 types of noise (filtered and unfiltered), 2 intensity levels, and 5 different test-tones upon the postexposure thresholds of 7 Ss.

Source of Variance	df	Sum of Squares	Mean Square	F
Filtered vs. Unfiltered Noise.....	1	1408.11	1408.11	142.96**
Intensity.....	1	584.25	584.25	59.31**
Frequency.....	4	337.85	84.46	8.57**
F. vs. Unf. \times Intensity.....	1	460.84	460.84	46.79**
F. vs. Unf. \times Frequency.....	4	347.18	86.79	10.14**
Intensity \times Frequency.....	4	157.57	39.39	4.60*
F. vs. Unf. \times I \times F.....	4	225.05	56.26	8.29**

B. $2 \times 5 \times 7$ Analysis of Variance of the 100 db Data

Source of Variance	df	Sum of Squares	Mean Square	F
Filtered vs. Unfiltered Noise.....	1	1740.00	1740.00	70.73**
Frequency.....	4	460.60	115.15	12.92**
F. vs. Unf. \times Frequency.....	4	548.70	137.10	22.77**
Conditions \times Ss.....	6	147.70	24.60	4.09*

C. $2 \times 5 \times 7$ Analysis of Variance of the 20 db Data

Source of Variance	df	Sum of Squares	Mean Square	F
Filtered vs. Unfiltered Noise	1	128.94	128.94	18.14**

* Significant beyond the .05 level of confidence.

** Significant beyond the .001 level of confidence.

Stimulation by unfiltered noise at 20 db SL resulted in significant fatigue only for the test-tones of 100 and 600 cps (Fig. 3). Stimulation by the filtered noise at 100 db SL resulted in statistically significant sensitization for the 162- and 255-cps test-tones, while fatigue was found for the 100- and 600-cps tones (Fig. 5). At the 20 db SL, statistically significant sensitization was found for only one test-tone, that of 350 cps. Signal detection was found to be significantly better after filtered noise stimulation than following unfiltered noise exposure.

Generally, the fatigue effect was more prominent and more consistent than the sensitization effect. The fatigue effects for the unfiltered noise were essentially the same for all the test-tones while the

sensitization effect following filtered-noise stimulation tended to follow the filtering effect. This may be seen by comparing Figs. 5 and 6.

Results indicate that fatigue and/or sensitization are greater following 100 db SL stimulation, than 20 db SL stimulation (Fig. 5).

The temporal course of postexposure thresholds was shown to vary as a function of whether there was initial poststimulatory fatigue or sensitization. Where poststimulatory fatigue was found, the fatigue tended to decrease with time (See Figs. 3 and 4). An analysis of the data of the individuals showed that where sensitization was observed initially, the next threshold was more sensitized in the records of four out of six Ss. In one of the

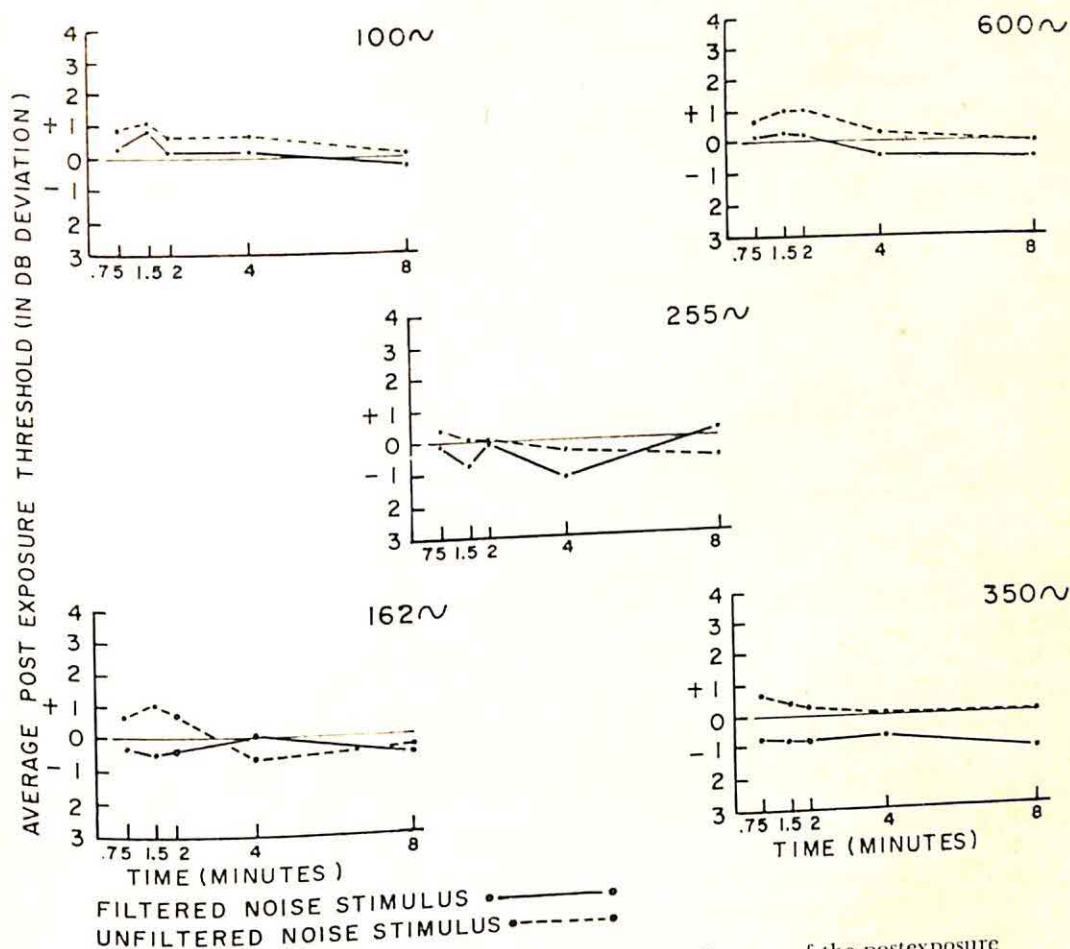


FIG. 3. The pooled data for all Ss, showing the temporal course of the postexposure thresholds for the 20 db SL stimulation.

two sensitization cases which did not respond as above, all postexposure thresholds were at essentially the same level. The other sensitization example showed that, after the initial sensitization, sensitivity decreased, increased, then decreased again. The temporal course of the postexposure thresholds is shown by Figs. 3 and 4.

During the course of the experiment four of the seven Ss consistently reported tinnitus following 100 db SL stimulation. At the end of the experiment the experimenter asked the S to match the pitch of a pure tone from a signal generator to the frequency of his tinnitus. The tinnitus following the unfiltered noise

was generally of a higher frequency than that following the filtered noise. In every case, the subjective tinnitus was of a higher pitch than the frequency of the highest test-tone used in the study.

V. DISCUSSION

It was found that thresholds for tones within the filtered gap were lower following filtered-noise stimulation and higher following unfiltered-noise stimulation. Hughes (15) explained sensitization as due to post-tetanic-potential. The results of the present study are opposite those that would be predicted on the basis of post-tetanic-potential. The post-tetanic-potential hypothesis

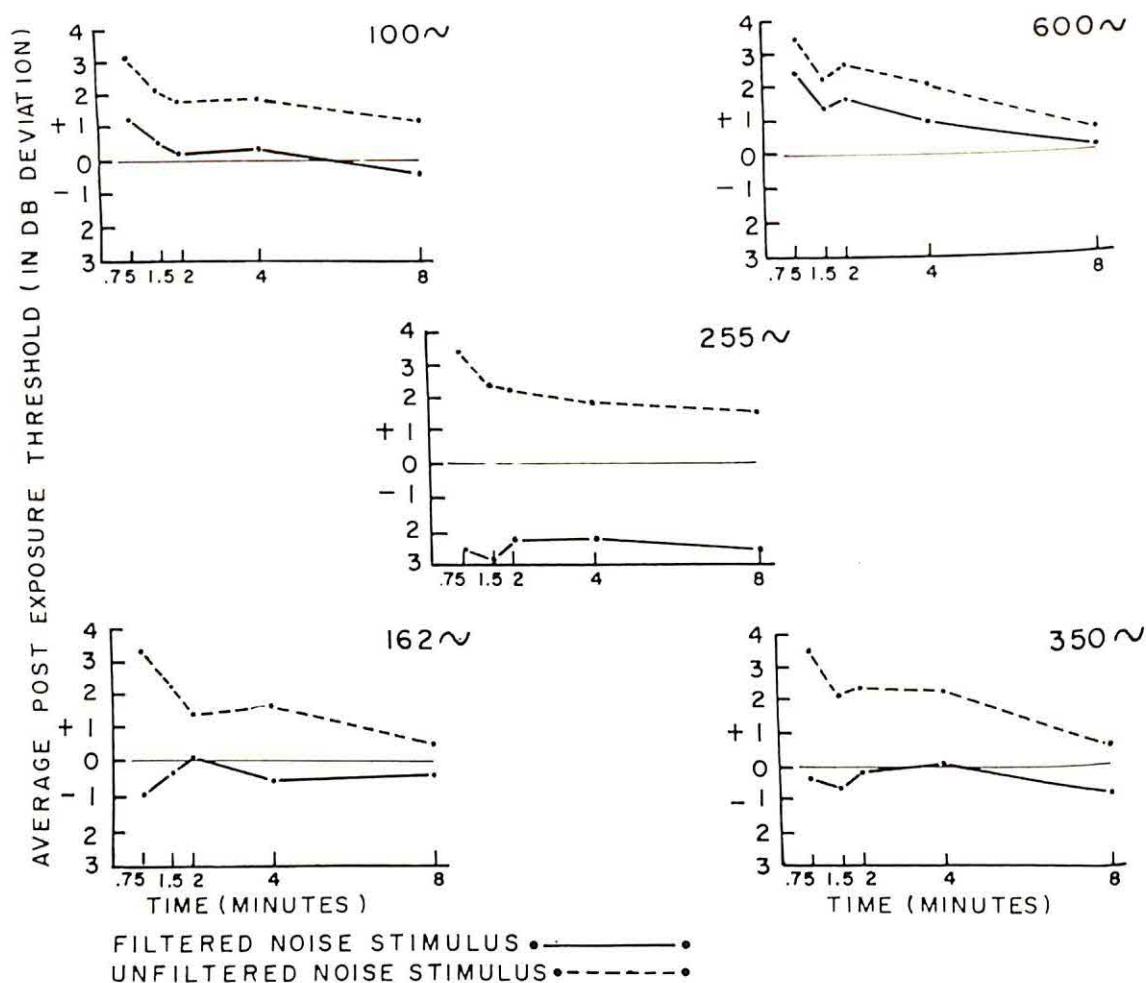


FIG. 4. The pooled data for all Ss, showing the temporal course of the postexposure thresholds for the 100 db SL stimulation.

would predict that more sensitization should result after the unfiltered noise.

The results of this study, though not explained in the literature, might be explained in a formulation borrowing from Tanner and Swets (29) their ideas regarding the hypothetical distribution of noise and signal-plus-noise for intensity, and their philosophy that the problem of detection is one of the detection of signals in noise.

It may be assumed that any sensory channel is continuously activated. The origin of this activity, hereafter referred to as ambient noise, may be ambient stimulation, neural noise, or both.

The ambient noise is assumed to approximate white noise. By definition, at a given frequency the intensities which would result if the noise were continuous would be distributed normally about a mean intensity which is the intensity of the noise. Assume further that the listener can discriminate perfectly (to be qualified later) among signals¹ of different frequencies, and in an experiment

¹ The term *signal* is used to refer to an experimentally produced stimulus while the word *stimulus* is used to mean any stimulation which activates the auditory system of the S. The term *stimulus* does, of course, include stimulation by ambient noise.

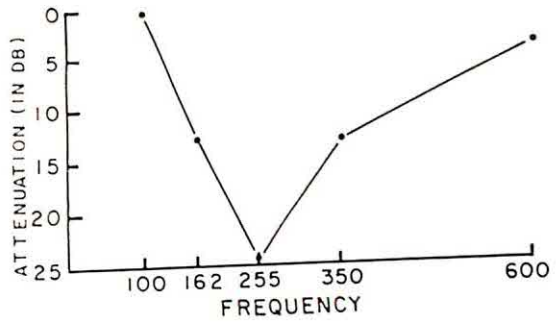
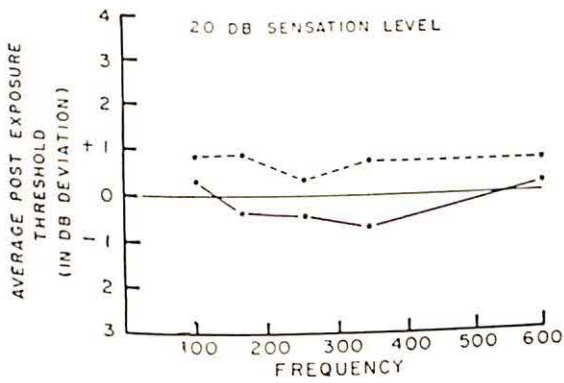


FIG. 6. Calibration curve for the filtered noise (on a linear scale to facilitate comparison with the fatigue and sensitization curves).

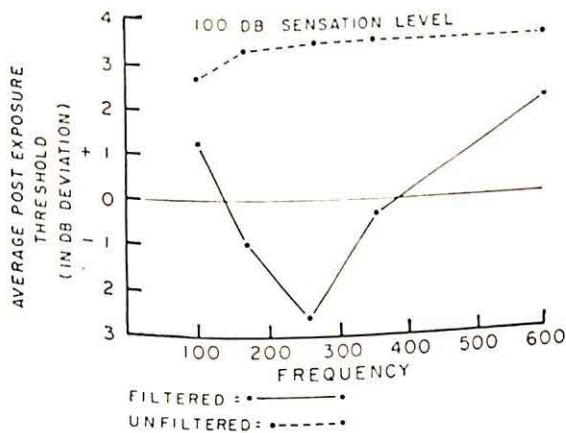
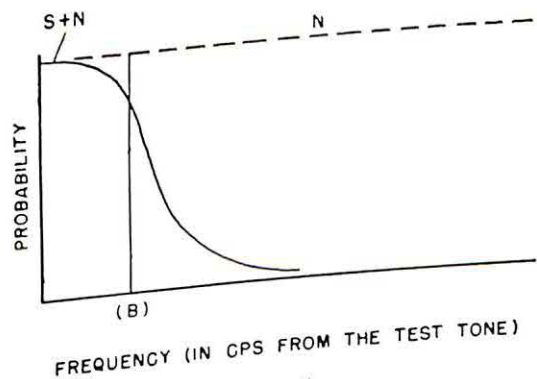
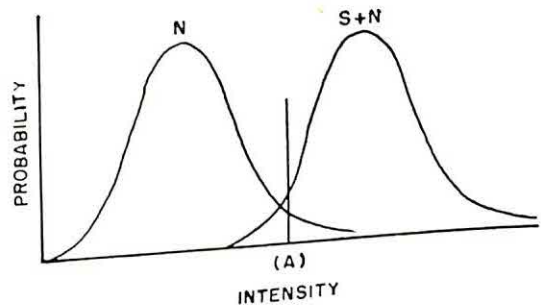


FIG. 5. Fatigue and sensitization curves for the filtered and unfiltered noise stimuli at two sensation levels. (This figure is derived from the pooled data of Figs. 3 and 4.)

moves the signal-plus-noise distribution farther out along the intensity axis. The noise distribution is unchanged by the addition of a signal; therefore, it becomes increasingly less likely that the *S* will say that more intense stimuli are coming from the ambient noise.

The following alternative exists. If



(A) INTENSITY CRITERION
(B) FREQUENCY CRITERION

FIG. 7. The distributions of noise (*N*) and signal-plus-noise (*S + N*) for intensity and frequency.

is set to detect signals of only one frequency, that of the test-tone.

The hypothetical noise and signal-plus-noise distribution for intensity and frequency may be seen in Fig. 7.

When an *S* is instructed to listen for a signal he attempts to decide whether or not a signal is present. *At all times* a stimulus of the expected frequency is present since it is a component of the ambient noise. Since intensities at a given frequency are normally distributed, when the *S* gets a stimulus of a given intensity, he must decide whether it is likely that the stimulus came from the noise that is always in the system. As the intensity of the stimulus increases it

the experimenter presents a signal of given intensity to the *S* the total distribution of intensities of a signal of that frequency changes, since the intensity distribution then has the ambient noise element as well as the signal component. If the signal is above the intensity level of the noise, the mean of the signal-plus-noise distribution will be higher than the mean of the noise distribution. In the absence of a signal, the *S* samples only the noise distribution. But when a signal is introduced, the *S* samples a signal-plus-noise distribution that differs from the noise distribution. Therefore, higher intensity stimuli are more likely from the signal-plus-noise distribution.

Although the *S* is instructed to respond in some way when he hears a specific signal, his actual task, if this explanation is correct, is to decide which of the two *distributions* he is sampling. The better guess which the *S* can make results in his responding "no signal" when he gets a stimulus that is closer to the mean of the noise distribution than to the mean of the signal-plus-noise distribution. Stated in different terms, the *S* establishes for himself an intensity criterion and any stimulus above this criterion will elicit "signal," while stimuli below this criterion will elicit "no signal."

If, as has been assumed, the two distributions are normal and have the same variance, the *S* would minimize his errors by setting his intensity criterion exactly halfway between the means of the two distributions. But it may be that he would not minimize errors in general, but minimize errors of a particular kind. If this were the case the intensity criterion would be set elsewhere. For example, if it were important to detect all signals but not important if false responses were made, the intensity criterion would be lowered.

Assuming that the *S* cannot discriminate perfectly among frequencies (or that the signal generator produces not only the desired frequency but also neighboring frequencies) and assuming (erroneously as has been previously shown) that intensity is constant, it is possible, by an argument similar to that above, to arrive at the following noise and signal-plus-noise distributions for frequency. Because of the physical nature of white noise, the distribution of frequencies is rectangular, i.e., all audio-frequencies are present and at any moment each is equally likely to appear. But the signal-plus-noise distribution for frequency is not flat. The signal generator is assumed to put out more signals at the test-tone frequency but it also puts out other signals that distribute themselves normally around the mean frequency which is that of the test-tone. It should also be noted that the *S* is brought in and exposed to his test-tone while his pre-exposure threshold is determined; therefore, he knows what signal to expect. But since the *S* cannot discriminate perfectly among frequencies he is more likely to respond "signal" to a tone of the same frequency as the test-tone and increasingly less likely to respond "signal" to tones as they differ in frequency from the test-tone. If this were true, the signal-plus-noise frequency distribution would be symmetrical about the modal point, i.e., the probability of a response to a stimulus would be decreased as the stimulus differs in frequency from the center frequency (which is that of the test-tone), regardless of whether the frequency difference is higher or lower with regard to the test-tone. The *S* sets some limit to the amount which the frequency of the stimulus may differ from that of the test-tone and still be called "signal." This may be called a frequency criterion. Thus, any stimulus within the frequency

criterion will be called "signal" and any outside this limit will be "no signal."

If the intensity and frequency assumptions made above are combined, omitting the erroneous assumptions of intensity and frequency constancy that were made for convenience, a three-dimensional graph of the probability of occurrence of a stimulus of given intensity and frequency can be drawn. Probability of a response can be determined once the probability of a stimulus' occurring is known. Figures 8 and 9 represent three-dimensional graphs of the probability of occurrence of a stimulus of given frequency and intensity.

Examination of Figures 7, 8, and 9 shows that part of the noise distribution is included within the region delimited by the *S*'s intensity and frequency criteria. This occurs since the two distributions (noise and signal-plus-noise) overlap and the *S*'s criterion would normally include some of the noise distribution. Also, not all of the signal-plus-noise distribution is included within the *S*'s criterion. This means that some noise stimuli will be called "signal" (a "false alarm") and some signals will be missed.

An additional assumption necessary in the above formulation is that auditory stimulation fatigues the *S* at the stimulating frequency and to neighboring frequencies in a gradient fashion with the maximum fatigue occurring at the test-tone frequency. Figures 8 and 9 show that stimulation at frequencies other than that of the test-tone would affect the sensitivity of the *S* in such a manner that the part of the signal-plus-noise distribution away from the test-tone would be maximally affected. The altered sensitivity of the *S* would also result in the noise distribution's being changed. The basic belief underlying the present formulation is that the relative volume of the noise and the signal-plus-noise dis-

tribution included within the area delimited by the intensity and frequency criteria determines auditory response. To elaborate, if an *S* is stimulated by a given signal, his sensitivity decreases for a band of stimuli. The stimulating frequency is at the center of the depressed band, the frequencies on either side of the center frequency are also depressed but to progressively lesser degree. If, after having been stimulated, the *S* has his threshold determined for a given tone for which his normal threshold is known, the change in sensitivity following stimulation can be observed. The above formulation holds that the stimulation would change the *S*'s sensitivity and therefore change the volume of the signal-plus-noise distribution relative to the noise distribution included within the criteria-passing region. If the postexposure signal were at the center frequency of the stimulated band, sensitivity would be maximally decreased. This would occur because the frequency signal-plus-noise distribution peaks at the signal frequency and drops off sharply thereafter. The noise distribution for frequency, on the other hand, is flat throughout its range. Therefore, a depression of the signal-plus-noise frequency distribution at the signal frequency removes relatively more signal-plus-noise than noise distribution. This change in the relative volume of noise and signal-plus-noise should be manifested as fatigue.

It is believed that essentially the opposite would occur when the *S* is stimulated by signals other than at the test-tone frequency. Stimulation by frequencies other than the test-tone frequency would affect the signal-plus-noise distribution in its less peaked area and would affect the noise distribution as before, in a flat area. Therefore, more noise and less signal-plus-noise volume

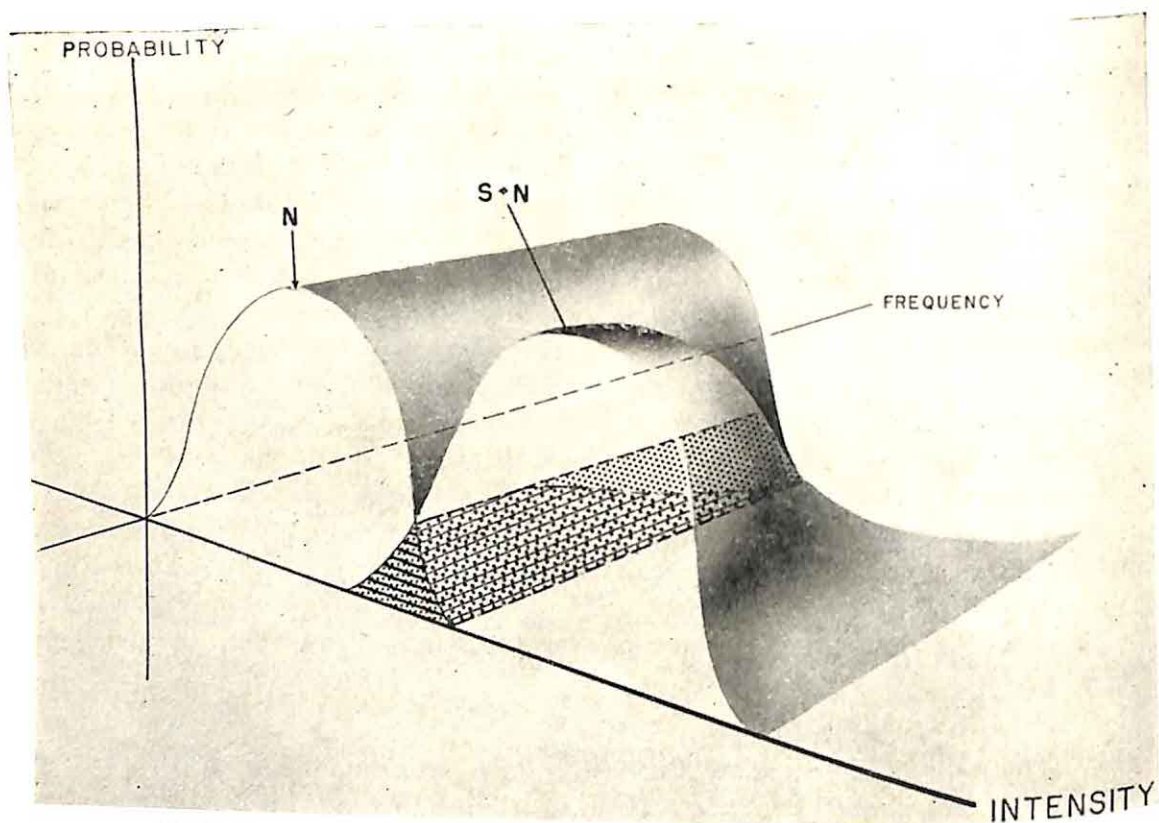


FIG. 8. A three-dimensional graph of the noise (N) and signal-plus-noise ($S+N$) distributions, showing the overlap between the two distributions.

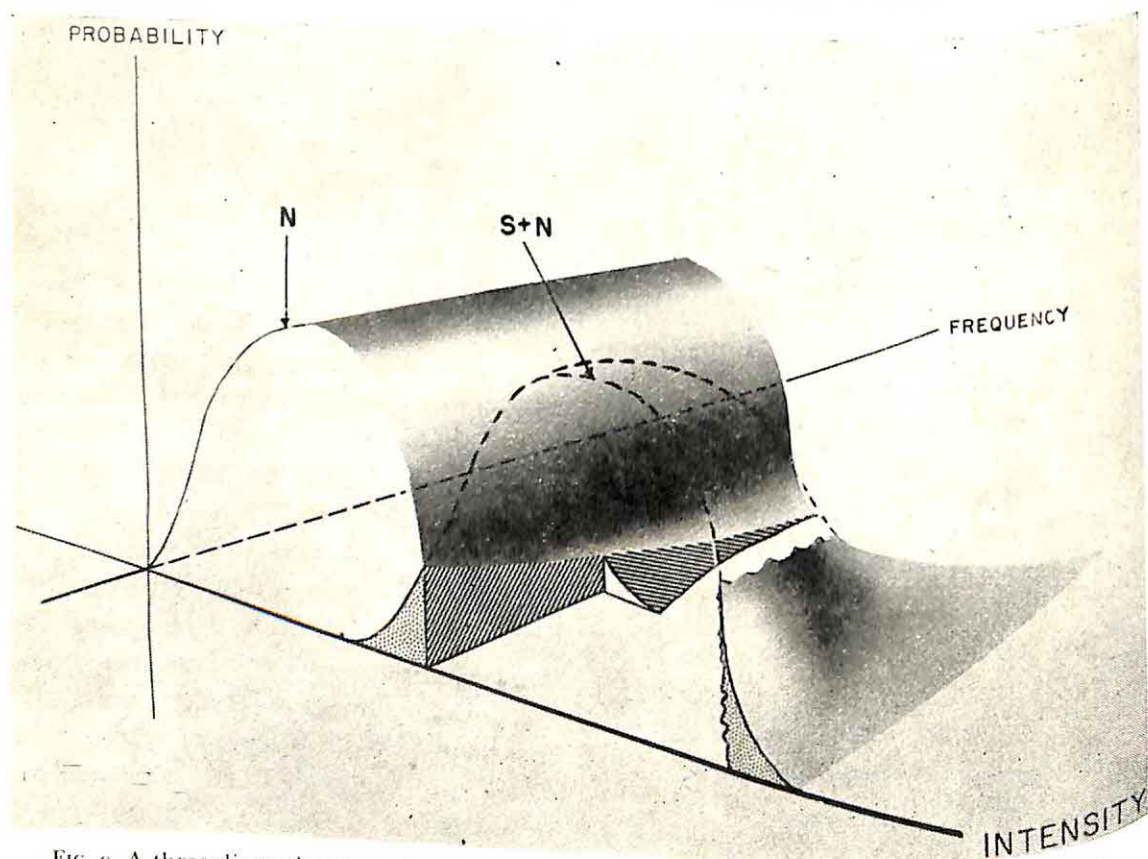


FIG. 9. A three-dimensional graph of the noise (N) and signal-plus-noise ($S+N$) distributions, showing the area delimited by the intensity and frequency criteria.

would be removed and sensitization would result.

However, the above formulation lacks sufficient generality to account for some of the previously cited sensitization results. Specifically, the above explanation predicts fatigue following stimulation at the test-tone frequency, while the previously cited studies report sensitization. It is apparent that the present formulation must be modified in order to account for these results.

The modification proposed is of the following nature: In this study and in that of Hughes (15) some Ss reported a postexposure tinnitus. This postexposure tinnitus could be assumed to approximate the frequency of the test tone and to change the shape of the noise distribution for frequency so that it is no longer flat, but peaks at or about the frequency of the tinnitus. If this occurred, whether fatigue or sensitization followed stimulation would be determined by the relative peakedness of the noise and signal-plus-noise frequency distributions. A more peaked noise than signal-plus-noise distribution would result in sensitization (following the previous argument) while if the noise distribution were less peaked, fatigue would ensue.

The significance of the main effects of intensity and frequency have been discussed. Unlike Hughes's study, sensitization following a low level of stimulation was found in the studies reported here.

The Hughes study reported a definite frequency-sensitization function such that test-tones and exposure frequencies below and above the range 300-500 cps produced different results from frequencies within that range. Examination of Figs. 5 and 6 show that in the present study the frequency-sensitization effects were quite similar to the filtering curve.

From this it is inferred that differential frequency effects in the range spanned by this study were small or nonexistent. If there had been noticeable differential frequency effects they should have made the frequency-sensitization curve deviate markedly from the filtering curve.

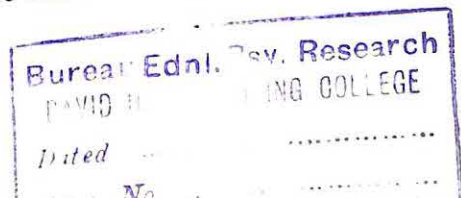
Figure 5 also reveals that fatigue effects for the different frequency test-tones were about equal. These results are consistent with those of Caussé and Chavassee (4), who report that in the frequency range of 100-600 cps spanned by test-tones in this study, little or no fatigue effects can be observed. Several further studies are suggested by the results of this study.

VI. SUMMARY

Pure-tone prestimulation thresholds were found for Ss who were then exposed for four minutes to the filtered or unfiltered noise stimuli at either a 20 or 100 db sensation level. A series of postexposure thresholds were then determined for the same tone for which the pre-exposure threshold had been found. The pre- and postexposure thresholds were then compared.

The Ss stimulated with unfiltered noise consistently had higher postexposure thresholds for pure tones within the filtered gap than did Ss exposed to the filtered noise. Following a 100 db SL stimulation by unfiltered noise, fatigue resulted for all test-tones. Following the 20 db SL only the 100- and 600-cps tones were fatigued. On the other hand, a 100 db SL stimulation by filtered noise produced sensitization for the 162- and 255-cps test-tones and fatigue for only the 100- and 600- cps tones. For the filtered noise at the 20 db SL, sensitization was found only at 350 cps.

Variations in poststimulatory sensitivity following the filtered noise were shown to be correlated with the relative



position of the test-tones with regard to the filtered gap. Results indicate that if fatigue follows 20 db SL stimulation, considerably more fatigue may be expected after 100 db SL stimulation. The same trend appears for sensitization.

The Ss were found to detect signals better after filtered-noise stimulation than after unfiltered-noise stimulation.

Within the limitations of this study the temporal course of postexposure thresholds was not found to be of a definite cyclical nature.

REFERENCES

- BRONSTEIN, A. J. Sensibilization of the auditory organ by acoustic stimuli. *Bull. biol. med. Exper.*, 1936, 1, 274-277.
- BUYTENDIJK, F. J., & MEESTERS, A. Duration and course of the auditory sensation. *Comm. Pontif. Acad. Sci., Rome*, 1942, 6, 557-576.
- CAUSSÉ, R., & CHAVASSE, P. Sensibilization auditive par stimulation binauriculaire discontinue. *C. R. Soc. Biol.* 1943, 137, 84-85.
- CAUSSÉ, R., & CHAVASSE, P. Etudes sur la auditive fatigue. *Année psychol.*, 1947, 43-44, 265-298.
- DAVIS, H., et al. Temporary deafness following exposure to loud tones and noise. *Final Report. Comm. Med. Res., OSRD Acta-otolaryng., Stockholm*, 1950, Suppl. 88.
- ECCLES, J. C. *The neurophysiological basis of mind*. London: Clarendon, 1953.
- EPSTEIN, A. A study of reversible auditory fatigue resulting from exposure to a pure tone. Unpublished doctor's dissertation, Univer. of Iowa, 1953.
- GEHARD, J. W. Motokawa's studies on the electric excitation of the human eye. *Psychol. Bull.*, 1953, 50, 73-111.
- GREISON, H. Comparative investigations of different auditory fatigue tests. *Acta-otolaryng., Stockholm*, 1951, 39, 132-135.
- HARRIS, J. D. Recovery curves and equinoxious exposure in reversible auditory fatigue following stimulation up to 140 db plus. *Laryngoscope*, 1953, 63, 660-673.
- HARRIS, J. D. The roles of sensation level and of sound pressure in producing reversible auditory fatigue. *Laryngoscope*, 1954, 64, 89-97.
- HARRIS, J. D., & RAWNSLEY, A. I. The locus of short duration auditory fatigue or "adaptation." *J. exp. Psychol.*, 1953, 46, 457-461.
- HARRIS, J. D., RAWNSLEY, A. I., & KELSEY, P. Studies in short duration auditory fatigue: I. Frequency differences as a function of intensity. *J. exp. Psychol.*, 1951, 42, 430-436.
- HIRSCH, I. J., & WARD, W. D. Recovery of the auditory threshold after strong acoustic stimulation. *J. acoust. Soc. Amer.*, 1952, 24, 131-141.
- HUGHES, J. R. Auditory sensitization. *J. acoust. Soc. Amer.*, 1954, 26, 1064-1070.
- JEFFRESS, L. A. Electronic switching circuit. *Amer. J. Psychol.*, 1950, 63, 257-259.
- JOSEPHSON, E. N. Auditory fatigue including a new theory of hearing based on experimental findings. *Ann. Otol., Rhinol. Laryngol.*, 1934, 43, 1103-1113.
- KOHLER, W. Zur theorie des sukzessivvergleichs und der zeitfehler. *Psychol. Forsch.*, 1923, 4, 115-175.
- LUSCHER, E., & ZWISLOCKI, J. Adaptation of the ear to sound stimuli. *J. acoust. Soc. Amer.*, 1949, 21, 135-139.
- MYERS, C. K., & HARRIS, J. D. Variability of the auditory threshold with time. *U. S. Naval Med. Res. Lab. Rept.*, 1950, No. 165, 9, 230-257.
- PATTIE, F. A. An experimental study of fatigue in the auditory mechanism. *Amer. J. Psychol.*, 1927, 38, 39-58.
- RAWDON-SMITH, A. F. Experimental deafness: further data upon the phenomenon of so-called auditory fatigue. *Brit. J. Psychol.*, 1936, 26, 233-244.
- RAWNSLEY, A. I., & HARRIS, J. D. Studies in short duration auditory fatigue: V. An investigation of the spread of fatigue within narrow frequency limits. *U. S. Naval Sub. Base*, 16 May 1952, Rep. No. 199.
- RAWNSLEY, A. I., & HARRIS, J. D. Studies in short duration auditory fatigue: II. Recovery time. *J. exp. Psychol.*, 1952, 43, 138-142.
- ROSENBLITH, W. A. Auditory masking and fatigue. *J. acoust. Soc. Amer.*, 1950, 22, 792-800.
- ROSENBLITH, W. A., GALAMBOS, R., & HIRSCH, I. J. The effect of exposure to loud tones upon animal and human responses to acoustic clicks. *Science*, 1950, 111, 569-571.
- ROSENZWEIG, M. R., & ROSENBLITH, W. A. Response to auditory stimuli at the cochlea and at the auditory cortex. *Psychol. Monogr.*, 1953, 67, No. 13 (Whole No. 363).
- SMITH, K. R. The problem of stimulation of deafness. II. Histological changes in the cochlea as a function of frequency. *J. exp. Psychol.*, 1947, 37, 304-317.
- TANNER, W. P., & SWETS, J. A. A decision-making theory of visual detection. *Psychol. Rev.*, 1954, 61, 401-409.
- WEBSTER, J. C., MILLER, P. H., THOMPSON,

- P. O., & DAVENPORT, E. W. The masking and pitch shifts of pure tones near abrupt changes in a thermal noise spectrum. *J. acoust. Soc. Amer.*, 1952, **24**, 147-152.
31. WEVER, E. G., & SMITH, K. R. The problem of stimulation deafness. I. Cochlear impairment as a function of tonal frequency. *J. exp. Psychol.*, 1944, **34**, 239-245.
32. WEVER, E. G. *Theory of hearing*. New York: J. Wiley, 1949.
33. WILCOXON, F. Some rapid approximate statistical procedures. New York: American Cyanamid Co., 1949.

(Accepted for publication April 30, 1956)

Effects of Light on Electrical Excitation of the Human Eye¹LORRIN A. RIGGS, JANET C. CORNSWEET,² and WARREN G. LEWIS³*Brown University*

THEORIES of color vision are based primarily on results of experiments on color mixture, spectral luminosity, effects of adaptation and contrast, and the wave-length discriminations of normal and color-blind subjects. These sources have yielded a large amount of information, but have failed to provide conclusive evidence on such basic matters as the number of fundamental response curves and the spectral characteristics of each.

In view of the failure of traditional approaches to the problems of color vision, it would seem desirable to follow any promising new leads that become available. Among these are the recording of gross electrical responses (electroretinograms) in the human eye when it is stimulated by lights of various wave lengths, and microelectrode studies of the responses of single retinal units in certain animal preparations. Unfortunately, the spectral sensitivity data obtained by Granit (7, 8) and others by microelectrode techniques of recording have not in the first place involved the human eye, and in the second place they have been based on somewhat indirect calculations. This follows inevitably from the fact that recording was from

third-order neurons, each of which was activated by a combination of numerous receptor cells. The human electroretinogram is even more a reflection of mass electrical activity; and while it has yielded some evidence for individual color response processes, it is far from a direct measure of the spectral characteristics of each.

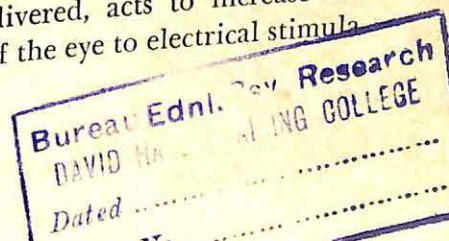
THE MOTOKAWA EXPERIMENTS

A new and different method of relating retinal phenomena to problems of color vision has recently been described by Motokawa (16, 17, 18, 19). This method is one in which the human eye is stimulated briefly by an electric current, using silver electrodes in contact with the skin of the brow and cheek. Currents of a few tenths of a milliamperere are sufficient to stimulate the retina. They give rise to "phosphenes," which present an appearance to the subject (S) of vaguely defined clouds of light in the peripheral field of view. An electrical threshold is determined by reducing the current in successive presentations of the test pulse until the S reports that the phosphene is no longer aroused. The electrical threshold so determined is found to be influenced by the state of adaptation of the eye and by other variables. Of particular interest to Motokawa is the observation that a flash of light, occurring before the electrical pulse is delivered, acts to increase the sensitivity of the eye to electrical stimuli.

¹ This investigation was conducted in the Psychological Laboratory of Brown University and was supported by a contract between the U.S. Office of Naval Research and Brown University.

² Now at the Department of Psychology, Yale University.

³ Now at the U.S. Medical Research Laboratory, New London, Connecticut.



tion. This suggests that the electrical threshold can be used as a measure of the excitability of the eye, so that the effects of light on the eye may be evaluated by measuring the alterations that the light produces in the electrical threshold. Specifically, Motokawa has undertaken a series of studies in which a flash of light has been followed at a given interval by a test pulse of direct current. A threshold is determined for the stimulating current, on the basis of its ability to arouse phosphenes in the stimulated eye. A comparison is made between the phosphene threshold following a light and the phosphene threshold in the dark. In this way a measure is obtained for the enhancement of the phosphenes by the preceding light. Since Gebhard (6) has reviewed these papers by Motokawa, no attempt is made here to give a complete account of them.

Of primary interest in Motokawa's work is the time course of the enhancement following lights of various colors. Figure 1 shows this time course for white, green, blue, and red lights. The ordinate is the zeta-value, Motokawa's measure of enhancement, where

$$\zeta = \frac{E - E_0}{E} \times 100$$

In this expression E is the momentary excitability or reciprocal of electrical threshold at some particular time after the flash of light, and E_0 is the reciprocal of an electrical threshold determined by delivering test pulses during a portion of the experiment in which no light is present. The abscissa is the delay time, i.e., the interval of time between the cessation of the flash and the onset of the electrical test pulse. In these experiments, the duration of the light is 2 sec. and the duration of the test pulse is 0.1 sec. It will be noted that ζ rises steeply immediately following the flash

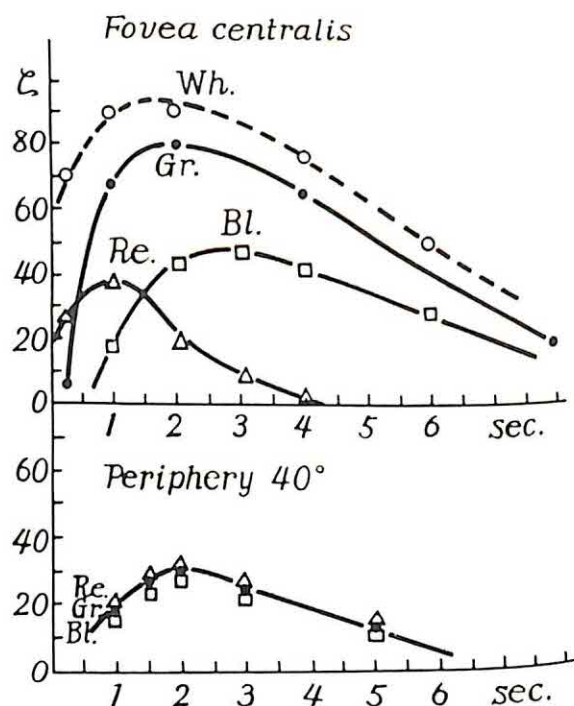


FIG. 1. Curves obtained by Motokawa (16) for the electrical excitability of the human eye as functions of time after flashes of light. Two-second flashes of red, green, blue, and white light were used with a 2° field centrally fixated (upper graph) and in the periphery (lower graph). The ordinate is the zeta-value as described in the text. These results of Motokawa are to be compared with Figures 6 and 9 of the present series of experiments.

of light. Thereafter, each function reaches a maximum at a delay time that is characteristic of the color of the preceding light. Motokawa reports that these "crest times" are highly stable at approximately 1 sec. for red, 2 sec. for green, and 3 sec. for blue light. The absolute size of ζ is related to the intensity of the light and varies somewhat from one S to another, but the shapes of the functions are said to be highly reliable and the crest time for each wave length exhibits little variation from one experiment to another. Hence Motokawa concludes that his method has revealed the specific temporal aftereffects of light in the three fundamental receptor systems for color in accordance with the trichromatic theory.

The curves are labeled R , G , and B in Fig. 2 because Motokawa believes that these delay time

The seemingly direct approach that Motokawa has used is based on certain assumptions whose validity may certainly be questioned. Implicit in the use of this method is the assumption that, at a delay time of 1 sec. after the flash, the electrical sensitivity of the eye is enhanced primarily by the aftereffects of the light on the red response mechanism of the eye. It is on the basis of this assumption that the curve in Fig. 2 for a 1-sec. delay time is designated as *R*, the curve for the red response mechanism. It is similarly assumed that for the 2-sec. curve, *G*, it is the green mechanism that is predominantly involved; and for the 3-sec. curve, *B*, the blue mechanism is assumed to predominate. It is difficult to understand how these assumptions can be made, however, in view of the shapes of the curves in Fig. 1. These curves indeed exhibit maxima at approximately the times indicated for red, green, and blue light respectively. But they also indicate that (a) at 2 sec. the effect of red light is fully half as great as at 1 sec.; (b) at 2 sec. the effect of blue light is almost as great as it is at its peak of 3 sec.; and (c) green light is almost equally effective in producing enhancement of electrical sensitivity at delay times of 1, 2, and 3 sec. after the flash. In other words, *the curves of Fig. 1 are not so sharply peaked at the times of*

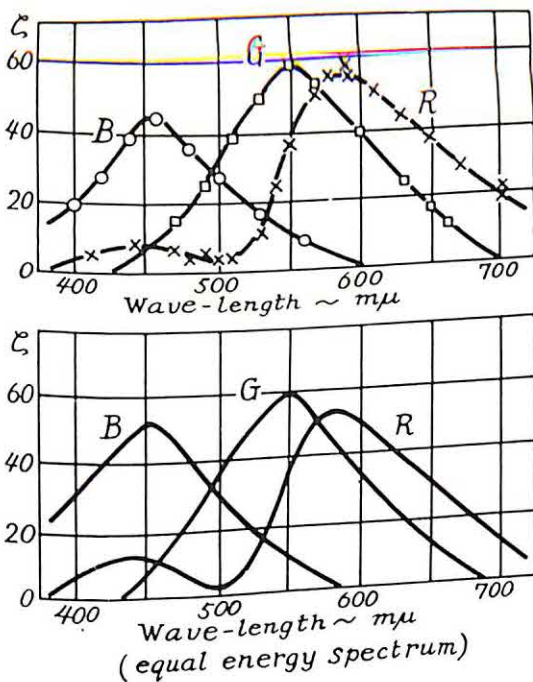


FIG. 2. Curves obtained by Motokawa (16) for the electrical excitability of the human eye as functions of wavelength of preceding flash of light. Two-second flashes, 2° central field. B, curve obtained for a 3 sec. delay time after the flash. G, curve for a 2 sec. delay time. R, curve for a 1 sec. delay time. The lower graph contains the three "physiological sensation curves" obtained by correcting the upper curves for spectral energy distribution.

1, 2, and 3 sec. as to suggest that each of these particular times is likely to characterize exclusively the delay time of any one of the (presumptive) three fundamental response mechanisms for color.

Motokawa has taken some account of the above difficulty by assuming that the degree of enhancement at a delay time of 3 sec., for example, is not exclusively determined by aftereffects occurring within the blue response mechanism. Enhancement is greatest, and therefore the threshold for electrical stimulation is lowest, with the blue mechanism at a 3-sec. delay time. But some degree of enhancement is thought to exist within the green mechanism and also the red. Hence the concept of "multiple thresholds," which plays an important part in Motokawa's theory of enhancement. It is Motokawa's contention that the lowest, or "true" threshold is one whose value at 3 sec. delay time is fixed by the action of the blue response mechanism. Higher thresholds exist, however; and these "apparent" thresholds result from the lesser actions of the green and red mechanisms at this delay time.

In summary, the work of Motokawa and his associates has offered a new approach to the problems of human color vision. Our interest has centered on the "physiological sensation curves" that have emerged from the experiments involving direct electrical stimulation of the eye. These curves have been worked out on the basis of test pulses of direct current applied to the eye at fixed intervals following flashes of colored light. The electrical test procedure reveals that electrical sensitivity is enhanced by light, and that the time course of the enhancement is specific to the wave length being used for the flash. For the extension of these basic experiments to the topics of fatigue, form perception, etc., the reader is referred to Gebhard's comprehensive summary (6) and the original works cited therein.

THE PURPOSE OF THE PRESENT INVESTIGATION

Our main goal in undertaking these experiments has been to furnish an in-

dependent check on the basic wave-length effects reported by Motokawa and his collaborators. We have made no attempt to repeat the whole extensive series of experiments. Instead we have chosen to concentrate on the basic phenomenon underlying much of this research, namely that of specific wave-length effects in the enhancement of electrical excitability following flashes of light. We have made every effort to observe wave-length effects under varying conditions of stimulus area and intensity.

The method followed in these experiments was fundamentally that developed by Motokawa. However, we felt compelled to introduce certain modifications in the interests of standardization of the procedure. These modifications, described more fully in later sections of this report, are considered important. They include the following: (a) Provisions for defining the intensity of electrical stimulation in units of current, independent of changes in skin resistance. (b) The use of equal step intervals in the descending series of stimuli for all threshold determinations. (c) Automatic control of all time intervals used in the presentation of light and current pulses. (d) A prearranged protocol for the presentation of the stimuli, for recording every judgment made by the S, and for the termination of the descending series as the criterion for threshold is reached. (e) Frequent determination of reference electrical thresholds without light in order to define the effects of light upon the threshold. (f) Employment of all the obtained data rather than discarding any results on the basis of variability in reference thresholds.

These modifications were introduced with the expectation that they would serve to reduce the effects of uncontrolled variables in such experiments.

For this reason it was anticipated that the effects of light itself would be more readily apparent in the results. It is apparent, however, that the use of these modifications means that we have not repeated exactly the experiments of Motokawa. Thus any basic differences between his results and ours are perhaps attributable in part to differences in procedure. Because of our concern with procedure we have made at the outset an extensive study of the variability characteristics of the basic method used by Motokawa as compared with the conventional method of constant stimuli.

APPARATUS AND PROCEDURE

In all of the experiments to be reported here, the *S* was seated in a light-tight, electrically shielded room. The experimenter was in an adjoining room. A thin partition separated the two rooms as indicated by the block diagram in Fig. 3.

Arrangements for Stimulation of the Eye

Optical system. The optical system is adapted from one described by Johnson (11) in an earlier report from this laboratory. It consists of

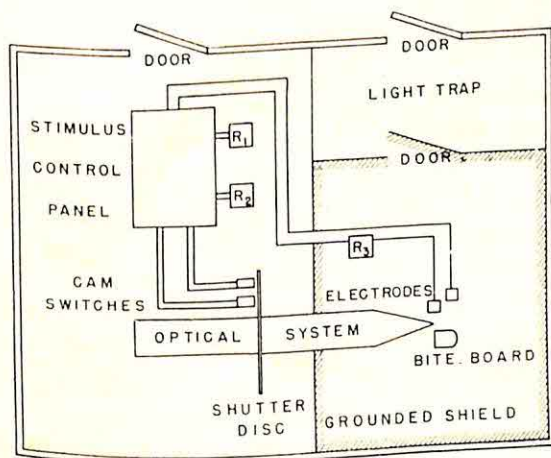


FIG. 3. Block diagram of experimental arrangements. Subject's room on right, experimenter's on left.

a tungsten source, collimating lenses, neutral density and color-selective filters, a rotating disc shutter, and a final lens that focuses the light on a spot within the pupil of the *S*'s right eye. Auxiliary devices include a dull red fixation point and a diaphragm to limit the size of field. A recent attachment to the apparatus is a special eyepiece that may be inserted to extend the field diameter to 38° . This consists of a negative lens, placed between the eye and the last lens of the usual optical system, and a positive lens of short focal length used next to the eye. The negative lens renders parallel the rays emerging from the regular optical system, and the positive lens focuses them on the pupil of the eye. A bite board of dental impression wax is used to position the head of the *S* so that no part of the light is occluded by the edges of the pupil. This positioning is easily accomplished by an experienced *S*, since the rays emerging from the apparatus are brought to a focus that is smaller than the natural pupil of the eye. The *S* views the last lens of the system filled with light ("Maxwellian view"). The fixation point appears to be at the center of this lens.

Calibrations of the intensities of the stimulating lights were made by the procedure outlined in an earlier report (Riggs, Berry, and Wayner, 23). By selecting appropriate filter combinations it was possible to present filtered light stimuli of various dominant wave lengths and of intensities such that they were of equal effectiveness in stimulating the eye. This equality could be achieved on the basis of either a photopic or a scotopic luminosity function.⁴

⁴ Because of the use of the Maxwellian view, the rays of light were not affected by the contraction of the natural pupil. Johnson (11) has discussed the resulting difficulty in specifying the photometric intensities of such stimulus fields. In the present experiment, the small ($2.8'$) field of white light reduced 3.7 log units by neutral density filters matches a comparison field of white light adjusted to 12.8 foot-lamberts and viewed with the natural pupil. Colored lights were provided by selective filters described by Riggs, Berry, and Wayner (23). On the basis of their calculations of the photopic equivalence of these stimuli we may state that each of the following combinations has the stimulating effect of a 12.8 ft.-L. field: White, with neutral density filters of 3.7 log units density; red (C) with 2.3 log units of neutral density filters; green (G) with 2.3 log units of neutral density filters; and blue-violet 76 with no neutral density filters. These stimuli are calculated to be roughly equivalent in their effectiveness for arousing the central cone receptor system of the retina.

The insertion of the special eyepiece (together with the removal of an aperture stop in the apparatus) has the effect of reducing the lumi-

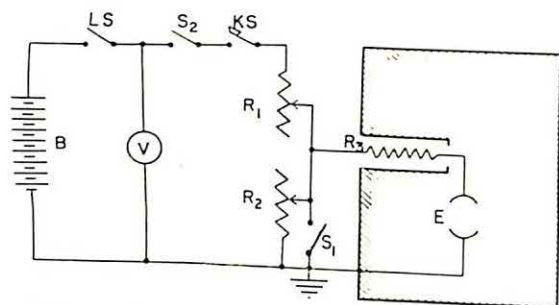


FIG. 4. Diagram of the circuit used for electrical stimulation of the eye. B, 90 volt battery; LS, line switch; V, voltmeter; S_1 S_2 cam operated switches; KS, experimenter's knife switch; R_1 , R_2 , resistance coils of potentiometer; R_3 , fixed resistance of 200,000 ohms; E, stimulating electrodes.

Timing of the optical stimuli was provided by means of a rotating disc shutter that intercepted the beam at its narrowest focus. An open section of the disc permitted an exposure of 2 sec. during each cycle of approximately 15.3 sec. The "wiping time" for opening or closing the shutter was .03 sec.

Electrical system. A diagram of the apparatus for stimulating the eye with electric current is shown in Fig. 4. A D.C. voltage is supplied by the battery at B in the diagram. This is nominally a 90-volt supply but its actual value is indicated by the voltmeter, V. This voltage is divided by the use of a potentiometer circuit containing two Cenco decade resistor units, R_1 and R_2 . The dials for these units were manipulated in such a way that the sum of the resistances set into R_1 and R_2 was always equal to 10,000 ohms. Thus, the voltage available for stimulating the eye was always proportional to the resistance set into R_2 ; specifically, this voltage, V_s , is given by the relation

$$V_s = \frac{VR_2}{10,000}$$

where V is the battery voltage indicated by the D.C. voltmeter. R_2 can be adjusted by fixed

nance while increasing the size of the field in order to include more of the peripheral rod-receptive region of the retina. Now white light with 2.0 log units of neutral density is the equivalent of 11.3 ft.-L. Again referring to the work of Riggs, Berry, and Wayner we find that there is an approximate scotopic equivalence of white light with 2.6 log units of neutral density filters; red (C) with no neutral filters; green (G) with 1.5 log units; and blue 76 with no filters. These stimuli are calculated to be roughly equivalent in their effectiveness for arousing the peripheral rod receptor system of the retina.

steps of 1 ohm from 0 to 10,000 ohms, but in actual operation steps of 10 ohms or more are used over a range of 1,000 to 7,000 ohms.

The electrodes at E consist of two silver plates held in contact with the skin of the brow and cheek at regions as near the eye as convenient directly above and below the eye. The cheek electrode has a square contact surface of 30×30 mm. It is grounded and serves as the cathode in electric stimulation. The electrode on the brow is a rectangle measuring 32 mm. long by 25 mm. high. The electrodes are held against the skin with approximately constant pressure by the use of a headband and spring device. Electrode paste is used to ensure good electrical contact with the skin. The electrical resistance is of the order of 4,000 ohms across these electrodes.

In series with the S is a fixed resistor, R_3 , of 200,000 ohms. This is a precision resistor entirely enclosed within a metal shield. Two important functions are served by this arrangement. First, it limits the current that can flow through the S so that even a charge of several hundred volts in the experimenter's room would be harmless to the S. Second, it permits an accurate statement of the current flowing through the S in terms of the setting of R_2 . The stimulating current, I_s , in amperes is given by the relation

$$I_s = \frac{V_s}{200,000 + R_s}$$

where V_s is the voltage supplied by the potentiometer as indicated above, and R_s is the resistance of the S. It will be noted that, with R_s typically about 4,000 ohms, even a relatively large change in R_s has very little effect on the value of I_s . It may therefore be said that, to an accuracy of one or two per cent, the stimulating current is proportional to the value of R_2 in the potentiometer, and its absolute value, I_s , may always be specified in terms of the above equation.

It has been found necessary to ground the stimulating circuit and place the S within an electrically shielded room. The cheek electrode was grounded, and the polarity of stimulation was such that the forehead electrode was made positive by reference to it.⁵

Timing arrangements. The switches S_1 and S_2 are microswitches actuated by a cam device mounted on the rotating disc shutter. The rela-

⁵ Motokawa has indicated that polarity has little effect on the degree of enhancement. We have confirmed this for the 0.1-sec. pulse duration used with the dark-adapted eye. Howarth's (10) experiments have shown that both on-and-off effects are found with this duration and that it is a suitable duration because the strength-duration curve has reached a steady value at about 0.1 sec.

tion between the open sector and the cam could be adjusted by reference to a radial scale, so that any desired delay time could be provided between the end of the light flash and the onset of the D.C. pulse. The cam operates in such a way that, just prior to the time at which a stimulus is to be given, S_1 is closed and then S_2 is closed. Then stimulation is begun when S_1 drops open, allowing current to flow through the S , and 0.1 sec. later S_2 drops open, thus terminating the flow of current from the potentiometer. These switches remain open until nearly the time for the next pulse, when they are closed in the order $S_1 \rightarrow S_2$ as before. While the full cycle of the shutter disc is 15.3 sec., it is possible to provide electrical pulses separated by only half of this interval by inserting a second cam in a position diametrically opposite to the first. This is ordinarily done for speeding up the determination of electrical thresholds in "dark" series, i.e., when no light flashes are involved. Control experiments have shown that similar thresholds are found with stimulus intervals of 15.3 and 7.65 sec.

The experimenter uses a noiseless knife switch, KS, to regulate the presence or absence of current, for those series in which the S must make a judgment of this kind.

Controls. The timing and intensity of electrical stimulation were continuously monitored by the use of a cathode ray oscilloscope. In addition, photographic records were made from time to time in which both the flash of light and the square-wave electrical pulse were registered on moving photographic paper. A Hathaway oscillograph was used for this purpose. This gave assurance that the various stimuli were being delivered at the proper times with the designated durations. It was found necessary to renew the microswitches from time to time. They typically delivered a clean square wave for several months of operation, but then developed a slight chatter or raggedness of action. Although the total energy of the stimulus was scarcely altered by this degree of irregularity, switches were discarded as soon as the raggedness became apparent on the oscillographic record.

Psychophysical Methods

Since one of the aims of the present series of experiments was to verify certain of the basic findings of Motokawa, it seemed important to use psychophysical procedures as similar as possible to those used in the original work. We have established a procedure for determining thresholds which incorporates the essential elements of the method which

he originated and uses in his work.⁶ Certain differences between this method, which will be called the limits-comparison (LC) method, and that of Motokawa will be discussed after the LC method is described.

The limits-comparison procedure. The LC method, as the name implies, is a two-stage procedure. It begins with a descending series of stimulus presentations, as in the conventional method of limits. The series is begun with the presentation of a stimulus which produces a clearly visible phosphene. The exact intensity of this first stimulus is varied from one determination to the next. After a positive response, the intensity of the stimulus is decreased to a value 98 per cent of the preceding one. This procedure is continued, with each electrical stimulus being 98 per cent of the preceding one, as long as the S continues to respond positively. When the S fails to see a phosphene, the same intensity is repeated on the next trial. If a phosphene is seen on the second presentation of the stimulus, the intensity is again lowered for the next trial. When the S responds negatively on two consecutive presentations of the same intensity, the second stage of the procedure is invoked.

The comparison stage of the method consists of presenting a pair of stimuli, one of which is a blank. The S reports which of the two is the test stimulus. The position of the blank, i.e., first or second, is randomized and established before the determination is begun. The S can respond by naming the position, first or second, at which he sees a phos-

⁶ We wish to express our appreciation to Dr. Motokawa for sending us descriptions of his method and for visiting our laboratory with the purpose of acquainting us with the details of his procedure. The most detailed published accounts of Motokawa's procedure for determining thresholds are found in references 20, 21, and 22.

phene, or, if he cannot discriminate, by saying "I don't know." The *S* is instructed not to guess, and to make a judgment only if he is able to state which position is correct. After each judgment of position is made the experimenter tells the *S* whether he is right or wrong. For those trials in which the *S* is unable to state which position is correct, the experimenter does not inform him of the correct position.

The comparison procedure is started by presenting a pair of stimuli, one of which is a blank and the other a test stimulus whose intensity the *S* has failed to see on two consecutive presentations in the method of limits. If the *S* correctly identifies the position of the stimulus, the intensity is again decreased by 2 per cent and another comparison pair is presented. If the *S* makes an error in naming the position of the phosphene, a second comparison at the same stimulus intensity is presented. If this comparison is correctly judged, a third comparison at the same intensity is presented. If the third is also correctly judged, the intensity is decreased for the next comparison. Each time that the *S* reports that he cannot name the correct position, the same comparison pair is repeated. The criterion for threshold is three consecutive "I don't know" responses at a given intensity, or two incorrect responses out of three comparison pairs at a given intensity. These criteria are not combined; that is, one or the other must be met. An *S* can respond incorrectly once, say "I don't know" once or twice, and still continue to the next step by responding correctly on two further trials. The intensity of the criterion value is called threshold.

A record is kept of all stimuli presented during a threshold determination together with the *S*'s response to each. The number of minutes after the beginning of the session at which the

TABLE 1
SAMPLE PROTOCOL ILLUSTRATING THE
LIMITS-COMPARISON METHOD
(Subject WGL 4/20/54—Dark Threshold)

Limits Stage		Comparison Stage		
R_2 Setting	Response	R_2 Setting	Correct Position	Response
4090	Yes	3200	2	R
4010	Yes	3140	2	R
3930	Yes	3080	2	R
3850	Yes			
3770	Yes	3020	2	?
3690	Yes	3020	1	?
		3020	1	R
3620	No			
3620	Yes	2960	1	R
3550	Yes	2900	2	?
3480	Yes	2900	2	R
3410	Yes			
		2840	2	W
3340	No	2840	1	R
3340	Yes	2840	1	?
		2840	2	R
3270	Yes			
		2780	1	W
3200	No	2780	1	?
3200	No	2780	1	?
		2780	1	?

Threshold is at an R_2 setting of 2780 ohms, giving a current of 116 microamperes.

determination is begun and ended are also indicated. Table 1 is a sample protocol for one threshold determination, showing the intensities presented (in resistance units) and the *S*'s responses.

The interval separating the stimulus presentations is kept constant by means of the timing device described in the apparatus section. This interval is fixed at one of two values, depending upon whether the threshold is being determined with or without a flash of light preceding the electrical pulse. For reference or "dark" threshold determinations the interval is 7.65 sec., while for "light" threshold determinations the interval is 15.3 sec. This doubling of the time interval is the only difference in the schedule for light and dark determinations. Control experiments have shown no difference in the precision or mean value of threshold determinations in the dark with each of the time intervals. The *S* is made aware that the stimulus is being presented by the sound of the switches being activated, as described earlier. No warning is given for the presentation of the light. The *S*, however, becomes familiar with the rhythm of the cycle and has no difficulty in being prepared for the light, or the phosphene. Once the thresh-

old determination is begun, no break or rest period is used. Stimuli are presented continuously until the threshold is reached. On the rare occasions when the *S* is unprepared for the stimulus because of a wink, cough, etc., he is allowed to ask for a repetition of that stimulus. He is not allowed to ask for a repetition of the stimulus for the sole purpose of re-evaluating it.

The time required to determine a threshold varies considerably from *S* to *S* and from one determination to another. The experimenter determines the starting point arbitrarily, attempting to estimate this value such that approximately eight to fifteen steps will be used in the limits method before the *S* begins the comparison stage. The initial presentation must be seen clearly by the *S*. If it is not, the determination is started again at a higher intensity. Ordinarily, at least four or five intensities are correctly judged in the comparison procedure before reaching threshold, but there is great variability in this. For a dark threshold, the total time taken varies from three to twelve minutes, with an average of about five minutes. When light is presented, this time is about twice as long.

Differences between the LC method and that of Motokawa. As stated earlier, an attempt was made to make the LC method as similar as possible to the method originated by Motokawa. The differences between these methods must be made explicit, however, so that the reader can consider these differences in evaluating the results.

Two differences between the methods appear to be of major importance. The first of these differences is in our use of equal log steps throughout all series of stimulus presentations. Motokawa's method is to use large steps of ten per cent or more in the beginning of a threshold determination (i.e., during the limits stage) and then to use small steps during the final portion of the comparison stage. Around threshold his steps are approximately one per cent differences. Motokawa does not establish the steps that he will use before the determination is begun.

Our own method has been to use a logarithmic series of predetermined stimulus intensities such that each successive step in the series is 2 per cent lower than the preceding one. Our reasons for insisting upon this mode of presentation are the following: (a) Results by the method of constant stimuli have shown that there is essentially a Gaussian probability relationship between the probability of seeing a phosphene and

the log of the stimulating current. It is therefore appropriate to use equal log steps as the basis for any descending scale of stimulus intensities. (b) The use of equal log steps achieves, in a systematic way, Motokawa's twin aims of avoiding the use of very lengthy series of stimuli and yet basing a threshold determination on steps of small linear magnitude. (c) The use of equal log steps has the further advantage that there is no massing of stimuli at any one level of intensity. The probability of finding a threshold at any given intensity level is raised by increasing the number of stimuli presented at that level. In other words, an equal-step method is a safeguard against experimenter bias in the determination of thresholds. This point is discussed more fully in the final section of this paper.

The second major difference, related to the first, is in the rigidity of the presentation of stimuli and the criteria for threshold. Motokawa's observers are allowed to ask for the comparison procedure at any time during the limits stage at which they consider the detection of a phosphene to be difficult. They are also allowed to ask for a repetition of a comparison pair if they are not sure of a response, or if they are "confident" of a response which proves incorrect. Another feature of Motokawa's method is that he continues to present comparison pairs after the *S* is no longer discriminating. This is done because of the possible existence of multiple thresholds, considered important by Motokawa. The *S* of multiple thresholds will be treated later in this paper. The threshold criterion is not rigid in Motokawa's method, and a threshold is decided upon after this continuation of the comparison series has led to a number of failures to discriminate. Again in the interests of low variability and standardization of the procedure, in the LC method all stimulus presentations are determined by the *S*'s response to the previous stimulus, and the rigid criterion outlined above is used for establishing the threshold.

Subjects

The reliable observation of an electrically aroused phosphene is a task which requires rather extensive training. As a consequence, all of the experiments to be reported have used a small number of trained *Ss*, usually three or four *Ss* per experiment. Most of the *Ss* took part in several of the experiments.

The *Ss* were laboratory personnel, including graduate students and two of the authors. All were tested by Ishihara

plates (5th ed.) and found to have normal color vision. Those Ss who customarily wore glasses were allowed to wear them during the experiments. Most of the Ss were familiar with the general purpose of the experiments, but every effort was made to prevent them from responding in terms of extraneous cues, forced choices, or preconceived notions. In judging single stimuli they were instructed to pay no attention to serial position but to call each as it appeared to them. In the case of paired stimulations they were asked to designate the one that appeared to be the stronger of the two, but were never forced to choose between two that appeared equal.

All Ss were trained for from eight to ten one-hour sessions prior to taking part in an experiment. This training consisted of practice in observing and reporting phosphenes, in distinguishing phosphenes from blank stimuli, in following the routine of the limits-comparison procedure, and in observing phosphenes after light flashes. Training was continued until all Ss arrived at a point where repeated thresholds under the "no-light" condition were reasonably stable. As training progressed the discrepancy between thresholds determined under similar conditions decreased markedly. By the end of the training period the second of two consecutively determined thresholds seldom differed by more than 15 per cent from the first. Only one S was not able to achieve this degree of stability and found it difficult to distinguish a test stimulus from a blank. He was not used in the experiments.

EVALUATION OF THE LIMITS-COMPARISON PROCEDURE

Experiment I. The Reliability of Threshold Determinations

The LC procedure has been used in the majority of our experiments for the following reasons: (a) It is basically the method developed and used by Motokawa in obtaining the specific wavelength effects in which we are primarily interested. (b) The procedure is relatively fast. A simple threshold may typi-

cally be determined by a single LC run of about five minutes' duration. A comparable determination by the constant stimulus method would take 15 to 20 minutes, and a similar time would be required for a typical set of ascending and descending series in the method of limits. The slower methods cannot well be used in sessions involving the ten or more threshold determinations that we have found necessary to use in the present experiments.

Since the LC procedure is a relatively new one, there is little information as yet on its reliability except for the reports of Motokawa and his associates. The numerous papers by Motokawa convey the impression of extraordinary reliability in the obtained electrical thresholds. Specifically, there is the statement that successive determinations of a reference threshold ordinarily agree within a few percentage points; a discrepancy as large as 10 per cent is relatively uncommon for these reference thresholds taken at various times during a given experimental session. A discrepancy as large as this is taken by Motokawa to mean that something is wrong with the results of that session, and they are discarded.

The high reliabilities reported by Motokawa for his limits-comparison data are certainly not typical of absolute threshold determinations in general. Agreement within 0.1 log unit (or about 20 per cent) is ordinarily considered quite satisfactory for individual determinations of absolute thresholds in such an experiment as the course of dark adaptation. Howarth (10), Clausen (5), and Gebhard (6) have stated that they were unable to achieve such a high degree of reliability. It therefore seemed desirable to make a detailed comparison of the reliabilities of the new LC method and the conventional method of

constant stimuli (CS). The design of the experiment is such that each experimental session provides data from both methods on the same Ss under the same experimental conditions. It must be realized, however, that the LC method as used here is not precisely the method originated by Motokawa. It has been modified as described above in the interests of greater objectivity.

Specifically, we have performed a balanced-order experiment in which half of the time was spent in determining absolute electrical thresholds by the LC method and half of the time by the conventional CS method. A sufficient number of replications were obtained so that conclusions could be drawn with regard to the variability of threshold determinations by each of the two methods and the absolute magnitudes of the thresholds obtained by each. Due account was taken of the fact that the CS method requires a much longer time for threshold determinations. No attempt was made to find thresholds following flashes of light in this portion of the experiments, because it was felt that reference thresholds in the dark were basic to all the findings reported by Motokawa and that the method could more conveniently be evaluated by the use of simple conditions.

Psychophysical procedure. Experiment I compares the LC method, as described above, with the conventional method of constant stimuli. The CS method was used as follows to determine the threshold for electrically induced "phosphenes." Seven stimulus values, separated by equal log steps, were used. Each intensity was judged 20 times. A random arrangement of the seven intensities was presented at the rate of one stimulation every 7.65 sec. After a total of 70 presentations, 10 at each intensity, a one-

minute rest period was given. The order was then repeated in the opposite direction, making a total of 140 presentations. Six blank stimuli were inserted into the order at random intervals and the S was informed of the blank after his judgment was given. This served to warn the S against false positive judgments.

The S's task was to make a judgment on the presence or absence of a phosphene at each stimulation. He called "yes" or "no" after each and his response was recorded. No comments were made by the experimenter except following the response to a blank stimulus, when the S was told that the stimulus had been blank. If the S missed a presentation by being unprepared for it, he was allowed to omit it and the stimulus was repeated at the end of the series.

The percentage of "yes" judgments at each stimulus intensity was calculated and plotted on a probability plot against the log of the stimulating current. A straight line was fitted by eye to the points, and the threshold was taken as the intersection of this line and the 50 per cent level on the probability scale. No threshold was considered as part of the data unless the percentage of "yes" judgments covered at least the range from 25 per cent to 75 per cent for the intensities used.

Plan of each Session. Each session of Experiment I consisted of two constant stimulus (CS) threshold determinations, each taking 20 minutes to obtain, and two 20-minute blocks of threshold determinations by the limits-comparison (LC) method. The LC blocks could not be exactly 20 minutes in duration, since the time required to obtain a threshold varies. In this case, several thresholds were obtained within a period of approximately 20 minutes. The number of thresholds determined in 20 minutes

Bureau Ednl. Res. Research
DAVID H. ... COLLEGE
Dated
Acce. No

varied from two to five. A one-minute rest followed the first and third 20-minute periods in the session; and a three-minute rest followed the second period. Preceding the first period there was a 15-minute period of dark adaptation. During the latter part of this dark-adaptation period one threshold was determined by the LC method. This threshold was not used as part of the data, but served as a "warm-up" for the S. It also gave the experimenter an indication of the level of sensitivity for that day, allowing him to select appropriate starting points for the later LC determinations.

Experimental design. Experiment I consisted of 16 of the sessions just described. These were held approximately three times a week and no two were ever held on a single day. The two methods of threshold determinations, LC and CS, were presented in an ABBA order on half the sessions and a BAAB order on the other half. The presentation of the two orders was balanced.

The same seven intensity values were used for a given S for each of the 32 thresholds determined by the CS method. The values were selected on the basis of preliminary work with each S, and were not changed after the experiment was under way. As a result, there were a few instances in which no threshold could be determined for a given period because of failure to satisfy the condition that the positive judgments cover at least the range from 25 per cent to 75 per cent over the series of intensities used.

Three Ss participated in this experiment. All had had previous training in observing phosphenes and in the routine of the LC method, as described in an earlier section. In addition, all had had practice with the CS method, during

which the range of intensities to be used in Experiment I was established. This range was .40 log unit (96 to 212 μ a) for LAR, .476 log unit (63 to 188 μ a) for SHH, and .476 log unit (84 to 251 μ a) for WGL.

Results. The data from Experiment I allow two types of evaluation of the LC method of determining phosphene thresholds. First, this method can be compared with a more conventional psychophysical procedure, the CS method. Second, the large number of threshold determinations by the LC method makes possible an internal analysis of the characteristics of this new psychophysical procedure.

A comparison of the limits-comparison and constant stimulus methods. It has been noted that the LC method requires less time for the determination of a single threshold than does the CS method. For the purpose of comparing the methods and balancing times, an equal amount of time during each session was spent on each method. Thus, there were four 20-minute periods in a single session, two spent on each of the two methods. Since a single period for the LC method allows time enough for the determination of from two to five thresholds, while only one was obtained in each period with the CS method, the means of the several LC thresholds in a given period are compared with the CS threshold value. However, an outstanding advantage of the LC method is its quickness, and the method is typically employed for single determinations of threshold. It is therefore necessary to compare single LC thresholds with CS thresholds. In the results tabulated below the comparisons of the LC and CS methods are made both for the mean LC threshold for a period and the first LC of a period.

TABLE 2

THRESHOLD VALUES OBTAINED BY THE LIMITS-COMPARISON PROCEDURE (LC) AND BY THE METHOD OF CONSTANT STIMULI (CS) DURING EXPERIMENT I

(Values in the body of the table are resistance settings of R_2 in ohms)

(Values in the body of the table are resistance settings of R_2 in Ohms)

Subject LAR

Day	LC		CS	CS	LC		Day	CS	LC		LC		CS
	1st	Mn			1st	Mn			1st	Mn	1st	Mn	
1	2840	3220	3700	3000	3030	3050	2	3440	3410	3410	3200	3500	4240
4	2840	2880	3490	3490	2060	2000	3	3600	2780	2870	2530	2820	3870
6	3200	3050	3700	3870	3200	3260	5	3250	2680	2750	2730	2700	3700
7	3080	2930	3730	3730	3340	2000	8	3840	3340	2990	3080	2930	3050
10	3480	3070	3320	—	2780	2700	9	3760	3020	2790	2730	2730	3390
11	3080	2750	3500	3630	3200	3170	12	3930	3200	3170	2780	2700	3440
13	2230	2600	3270	3410	2000	3030	14	3400	3550	3680	2000	3090	3620
16	2580	2380	3140	3390	2780	2730	15	3570	3340	3150	2960	3020	3670
Mean	2916	2860	3402	3644	3136	3091		3610	3165	3101	2864	2947	3697

Subject SHH

Day	LC		CS	CS	LC		Day	CS	LC		LC		CS
	1st	Mn			1st	Mn			1st	Mn	1st	Mn	
2	2330	2230	2450	2940	2630	2530	1	2740	2780	2740	3020	2960	3020
3	2330	2400	2540	2540	2480	2560	4	2360	2330	2240	2430	2190	2660
5	2280	2100	2360	2420	2840	2950	6	2600	2630	2650	2680	2660	2600
8	2030	2280	2300	2540	2730	2600	7	2580	2630	2930	3080	3100	—
9	2330	2430	2520	2580	2110	2220	10	2390	2530	2360	2480	2430	2470
12	2580	2480	2430	2490	2330	2670	11	2500	2430	2380	2530	2460	2410
14	2030	2260	—	2400	2580	2480	13	2350	3270	3200	2780	2780	2670
15	2280	2530	2470	2570	2780	2760	16	2580	2680	2880	3270	2920	2370
Mean	2274	2339	2451	2560	2560	2596		2512	2660	2672	2784	2687	2600

Subject WGL

Day	LC		CS	CS	LC		Day	CS	LC		LC		CS
	1st	Mn			1st	Mn			1st	Mn	1st	Mn	
2	3340	3240	3100	3500	3140	3380	1	3130	3770	3270	2580	2430	2960
3	2680	2960	2880	2660	2780	2860	4	2600	3080	2760	3410	3130	2750
5	2630	2560	—	—	3770	2880	6	2980	3080	2940	2580	2510	2900
8	3020	2900	2960	3340	2630	2460	7	3120	2840	3200	2580	2790	3240
9	3550	3370	3040	3040	2900	2990	10	3340	3410	3840	3480	3430	3370
12	3410	3410	2870	3370	2900	2780	11	3310	3270	3340	3690	3450	3570
14	2780	2840	3070	2840	2840	2680	13	3140	2530	2320	2110	2090	3070
15	3270	2960	3010	2830	2730	2900	16	3100	3200	2770	2840	2820	2930
Mean	3085	3030	2990	3083	2961	2866		3106	3147	3055	2909	2831	3099

Absolute thresholds obtained by the two methods. Table 2 presents the CS threshold, the mean of the LC thresholds, and the first LC threshold for each of the four periods of each of the 16 sessions for each of the three Ss.⁷ Figure 5 shows the means of these val-

ues. It will be noted that each value shown in Fig. 5 is the mean of eight

⁷ The missing values for CS thresholds result from a failure to meet the criterion of basing a threshold on percentages of positive judgments running from 25 per cent or less to 75 per cent or more.

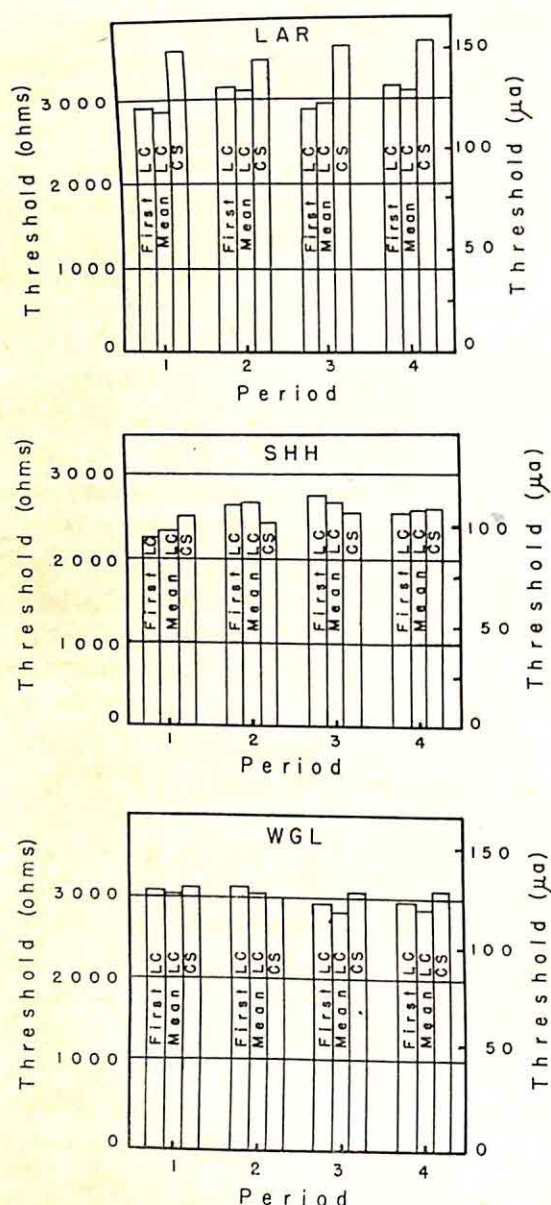


FIG. 5. Mean thresholds for the elicitation of phosphenes in the dark-adapted eye. Threshold values for the first LC, mean LC, and CS determinations in the four periods of Experiment I. The individual thresholds and their means are tabulated in Table 2.

sessions. The CS method was used eight times in each period in the 16 sessions, and the same is true for the LC method. Thus, in the bar diagrams of Fig. 5, Periods 1 and 4 for a given method represent the same eight sessions of the experiment, while Periods 2 and 3 repre-

TABLE 3

OVER-ALL MEAN THRESHOLD VALUES

(Values of R_2 in ohms for all periods combined for each method)

Subject	CS	LC(mean)	LC(first)
SHH	2531	2574	2560
WGL	3070	2946	3026
LAR	3610	3000	3020

sent the other eight sessions.

It can be seen from Fig. 5 that there is no consistent tendency for the threshold to increase as the session progresses for either method. It must be remembered that the Ss were dark-adapted for 15 minutes prior to the first period of a session. Hence the rapid changes in threshold that are found during dark adaptation (Achelis and Merkulow, 1; Lewis, 12) are not of any consequence here.

A second finding evident from Fig. 5 is the similarity of absolute threshold values for the two methods as evidenced by two of the Ss, SHH and WGL. The other S, LAR, had a consistently higher threshold when measured by the CS method. The over-all means for the three measures are presented in Table 3.

Variability of thresholds obtained by the two methods. Experiment I was designed so that the two methods (LC and CS) could be compared with regard to their within-session variability. Thus, each session consisted of two periods on each of the methods. Further, for half of the sessions a given method was used to determine thresholds in adjacent time periods (Periods 2 and 3) and for the other half the same method was used in separated time periods (1 and 4). Hence, the within-session variability of the methods can be compared for each set of time conditions. The day-to-day variability of electrical threshold determinations is large, and experiments on

this phenomenon are best designed so that thresholds taken on different days are not directly compared. Therefore, this type of variability will not be considered here.

The basic measure that will be considered here is the average within-session change for each method. The measure is obtained by computing the difference between the two thresholds determined by the CS method, for example, in a given session. Since the methods were presented in an ABBA order on half the sessions and a BAAB order for the other half, there will be, for each S, eight differences between Periods 2 and 3 for the CS method and eight differences between Periods 4 and 1. The mean of these eight differences gives a measure of the average change, or within-session variability for that method. An example of the computation of this average change is given in Table 4.

The differences in Table 4 were obtained by subtracting the threshold obtained in Period 1 from the threshold obtained in Period 4 for each of the sessions in which the CS method appeared in these periods. Reference to Table 2 in which these thresholds are

presented may clarify the procedure. The mean of these values is obtained, both with and without regard to the sign of the difference. A positive difference indicates that the threshold was higher, i.e., less sensitive in the later period. The algebraic mean is used as the basis for the calculation of the variance of the measure.

The above tabulation was followed for both the CS and the LC methods. For the LC method, it was done twice, once for the mean LC of the period and once for the first LC of the period. It was done for both the adjacent periods (3-2) and separated periods (4-1). Two main comparisons emerge from this tabulation. First, the absolute size of these within-session differences is of interest. The size typical of these differences is an indication of the variability of thresholds determined by each method. Second, the algebraic mean of these differences is also important, since it provides for the assessment of trends through the session. Positive mean differences would indicate a decrease in sensitivity as the session progresses. The algebraic mean is also the proper basis for a variance measure. Knowing the variance, one can make a statement of the limits of confidence of a given difference. Tables 5 and 6 present the results of this analysis of the two methods. In Table 5 the algebraic and absolute mean changes are presented, together with the over-all mean threshold values for each S. The mean thresholds, combining all periods, are presented in order to give meaning to the magnitude of the differences presented. Table 6 presents *F* ratios, comparing the variances of the methods.

A consideration of Table 5 shows that the CS method has generally given more reliable thresholds than has the LC

TABLE 4
SAMPLE COMPUTATION OF THE AVERAGE
CHANGE IN THRESHOLD VALUES
(Subject WGL-CS Method; Period 4-Period 1)

Day	Difference
1	-170
4	+60
6	-80
7	+120
10	+30
11	+260
13	-70
16	-170
Mean (algebraic)	- 2.5
Variance	22221.
Mean (absolute)	120

TABLE 5

AVERAGE CHANGES (OHMS, ALGEBRAIC AND ABSOLUTE) OF THRESHOLDS OBTAINED IN A SINGLE SESSION OF EXPERIMENT I BY A SINGLE METHOD

	LAR		SHH		WGL	
	Absol.	Algeb.	Absol.	Algeb.	Absol.	Algeb.
CS Method						
Period 3-2	127.14	+127.14	131.43	+131.43	272.86	+02.86
Period 4-1	350.00	+87.50	182.86	+07.14	120.00	-2.50
Mean Threshold	3610		2531		3070	
LC Method (mean)						
Period 3-2	176.25	-153.57	132.50	+15.00	356.25	-223.75
Period 4-1	323.75	+231.25	310.00	+257.50	278.75	-103.75
Mean Threshold	3000		2574		2046	
LC Method (1st)						
Period 3-2	313.75	-301.25	258.75	+123.75	443.75	-238.75
Period 4-1	395.00	+220.00	403.75	+286.25	448.75	-123.75
Mean Threshold	3020		2560		3026	

method. This conclusion is based on the average discrepancy between thresholds determined in the two periods of each session devoted to a given method. The average discrepancy for the CS method is lower than that obtained with the LC method even when due allowance is made for the fact that individual LC runs take less time than the CS determinations. The magnitude of the discrepancy relative to the size of threshold is in every case less than 10 per cent for the CS method, less than 15 per cent for the LC mean, and less than 20 per cent for the first LC. The average discrepancy between the adjacent periods, three and two, is consistently smaller than that of

Periods 4-1 for LAR and SHH, but not so for WGL. These 3-2 period changes are of particular interest, since they indicate the variability of thresholds determined closely in time, as in the typical experimental situation. Since they are generally smaller than those determined far apart in time, it is concluded that reference determinations should be repeated frequently.

The algebraic differences, also shown in Table 5, are not consistently positive; in other words, the threshold does not appear to exhibit a rising trend during the session. One of the Ss, SHH, does show some tendency toward a rise in threshold. Another S, WGL, shows the

TABLE 6
F RATIOS OF VARIANCE OF AVERAGE CHANGES

	LAR		SHH		WGL	
	LC _{mn} /CS	LC _{1st} /CS	LC _{mn} /CS	LC _{1st} /CS	LC _{mn} /CS	LC _{1st} /CS
Period 3-2	3.494	2.853	1.411	3.961	1.407	2.559
Period 4-1	1.672†	1.353	1.925	3.007	4.382*	15.153**

* $p < .05$.** $p < .01$.† In this case the CS variance was greater t' in the LC variance.

opposite trend as indicated by negative values in all but one of his discrepancies.

The F ratios shown in Table 6 point out the fact that in all but one case the variance of the LC method changes was greater than that of the CS. Only for LAR, 4-1 period difference, was the CS more variable than the LC mean but not more variable than the LC first. None of these ratios for the 3-2 period difference is significant. For the Period 4-1 difference, they are significant for WGL, again indicating that frequent determinations of a reference threshold are desirable. Variances combining the 4-1 and the 3-2 periods were also computed for both methods. Here the LC difference variance was always greater than the CS. None of these was significantly greater when the mean LC was used, but for two Ss, WGL and SHH, the LC first/CS ratio was significant. When the variance of the first LC is compared to that of the mean LC, no significant differences are found for any S. In all but one case, however, the first LC variance was greater than the mean LC.

Characteristics of the Limits-Comparison Procedure

Variability. The large number of repetitions of the limits-comparison method (two to five in each 20-minute period) makes it possible to compute the variability of thresholds determined by this method. Since the method involves presenting equal log steps, each representing approximately a 2 per cent decrement in stimulus intensity, the difference between any two thresholds may be expressed as the number of these steps separating the thresholds. It is convenient to make use of these step units for the purpose of setting up a distribution of differences between adjacent thresholds.

The details of the procedure which has been used to get at the question of variability are as follows. In each period there are several thresholds obtained by the LC method. The differences between successive thresholds within each period are found. These differences between the first and second, second and third, etc., are tallied for each S. The result is a distribution of differences, the mean of which represents the mean discrepancy between one threshold value and the next. The standard deviation of this distribution can be used to estimate fiducial limits of this discrepancy. As in the earlier computation of differences, this mean difference has been computed algebraically, to assess trends and to compute the variance, and absolutely, to estimate the average size of the discrepancy independent of direction.

The results of this analysis are seen in Table 7. The discrepancies in this table are expressed in step units.

A positive value in the algebraic mean data of Table 7 indicates that the second threshold was higher, i.e., less sensitive, than the first. The values labeled transition threshold will be considered in the next section. It may be noted that for

TABLE 7
AVERAGE ADJACENT THRESHOLD CHANGE
(in 2 per cent step units)
LC Method

	Abso- lute Mean	Alge- braic Mean	N	SD
LAR				
LC Threshold	4.77	-.46	71	5.84
Transition Threshold	3.61	-.06	71	4.49
SHH				
LC Threshold	3.41	-.03	59	5.11
Transition Threshold	3.95	+.15	59	4.96
WGL				
LC Threshold	5.10	-1.01	69	6.53
Transition Threshold	5.19	-1.30	69	8.15

each *S* the LC threshold difference is negative, but not large. From the information in Table 7, it can be predicted that the second of two adjacent thresholds will fall, 68 per cent of the time, between the limits of 6.30 steps below the first to 5.38 steps above the first for LAR, 5.14 steps below to 5.08 steps above for SHH, and 7.54 steps below to 5.52 steps above for WGL. The above values are obtained by taking one standard deviation above and below the mean discrepancy. Since each step represents approximately a 2 per cent change, the percentage discrepancy expected can be computed approximately by multiplying the above values by two.

Limits method vs. LC method. Since the first stage of the LC method is the descending phase of the conventional method of limits, a question arises as to whether the reliability of the method is increased by adding the comparison stage of the method. Motokawa reports that the method of limits by itself, without a comparison phase, is not sufficiently sensitive to be applicable to the study of electrical thresholds.

Although the present experiment did not include a direct comparison of the LC method with the method of limits, it is possible to make a comparison between threshold values obtained with the LC method and the transition intensity, i.e., the value at which the limits stage of the method was terminated by the occurrence of two successive judgments of "No." It should be pointed out that this value, which may be called the "transition threshold," is not actually equivalent to a threshold obtained by the usual method of limits, since the *S* was aware that the second, or comparison, stage was to follow. However, an analysis of these transition thresholds can serve as an estimate of the variability

which would be expected from the method of limits. Clausen (5) has compared the comparison method with the method of limits, using both ascending and descending series for the latter. While it is not clear that he made any allowance for the differences in time taken by the two methods, he concludes that the limits method yields threshold values higher but less variable than those obtained by the comparison procedure. It is clear that the transition threshold must always be higher than the LC threshold. However, the variability of the transition threshold can be compared with that of the LC threshold. The procedure followed to obtain the adjacent threshold discrepancies for the transition thresholds is exactly as described for the LC method thresholds. Table 7 includes these values. It will be noted that the changes obtained when the transition threshold is used are very similar to those obtained by the LC method, and that no consistent trend occurs in any of the three *Ss*.

We may conclude that the comparison phase of the LC method has failed to contribute to the reliability of thresholds obtained by the limits phase alone. The LC method was nevertheless used throughout the major portion of our experiments for the following reasons: (a) It is the basic method used by Motokawa to study specific wave-length effects, (b) it yields thresholds that are lower, and hence presumably more sensitive to the aftereffects of light, than those obtained with the limits procedure alone, and (c) the method has the advantage of maintaining the *S's* interest by informing him of the correctness or incorrectness of his comparison judgments.

Characteristics of the constant stimulus method. The data obtained by the CS method were typical of absolute

threshold data in general. The results of each threshold determination were plotted on probability paper, with percentage of "yes" responses on the probability ordinate and log stimulating current in microamperes on the abscissa. The resulting points showed the usual degree of conformity to a linear relationship. A straight line was fitted by eye to the points of each graph. There were 32 graphs for each *S*, two on each of the 16 sessions.

One index of the precision of the judgments is given by the ratio between the intensity that gives 75 per cent judgments of "yes" and the intensity that gives 25 per cent judgments of "yes." The mean value of this ratio is 1.42 (.152 log unit) for *S* LAR, 1.75 (.244 log unit) for *S* WGL, and 1.53 (.185 log unit) for *S* SHH. The standard deviations of the distributions of these values, expressed in log units, are 0.027, 0.053, and 0.049 respectively.

During the course of each threshold determination six blank stimuli were presented. That means that a total of 192 blanks were presented during the 32 determinations for each *S*. The numbers of false positive responses obtained with these stimuli are as follows: For *S* LAR, one (0.5 per cent); for WGL, 28 (14 per cent); and for SHH, 20 (10 per cent).

Conclusions with Regard to the Limits-Comparison Procedure

A conservative conclusion from the results of Experiment I is that the limits-comparison method has not proven to be more reliable than the conventional method of constant stimuli in the determination of thresholds for the appearance of phosphenes in response to electrical stimulation. Furthermore, the over-all precision that we have found for the electrical thresholds is not so

high in any of these experiments as in the ones reported by Motokawa. In the final "Discussion" section of this paper we have attributed the phenomenally low variability in Motokawa's data to the use of psychophysical procedures that serve to curtail the range of intensities within which a given threshold is likely to be found.

APPLICATION OF THE LIMITS-COMPARISON PROCEDURE TO THE STUDY OF ENHANCEMENT

General Procedure

Electrical stimulation experiments are of interest chiefly in the possible interactions between electrical and photic stimuli. Motokawa has reported that the aftereffect of a flash of light is nearly always to enhance the sensitivity of the eye to electrical stimulation. The remainder of the present experiments consists mainly of determining the amount and time course of this enhancement following single flashes of light. The procedure is first to find the basic sensitivity of the eye by determining the absolute threshold for electrical stimulation in the dark. Then successive flashes of light are presented, and the corresponding threshold is determined for electrical stimuli presented at a given delay time after each flash. The degree of enhancement is then expressed as the ratio of electrical sensitivities determined with and without the flash of light. This has been done for a number of different experimental conditions with regard to delay time, intensity of light, area of light, wave length of light, and certain procedural variations. The basic procedure is described below. In all the experiments dealing with the effects of light upon sensitivity to electrical stimulation, the LC method has been employed.

A single experimental session. The

experimental session is begun with the positioning of the S's biting board so that the light enters the pupil of the right eye. This is accomplished by turning on the light and allowing the S to make adjustments in the position of the biting board. The S learns to do this by looking successively up, down, to the right, and to the left, meanwhile adjusting the biting board so that the light disappears at about the same degree of rotation for each quadrant. After this is accomplished, the light is turned off and the S is dark-adapted for 20 minutes. During the latter part of this dark-adaptation period one threshold is determined by the LC method, under the no-light condition. This threshold is not used as a part of the data but serves merely as a practice determination.

After the S has been in the dark for approximately 20 minutes, the first dark, or reference, threshold is determined.⁸ Following this, a "light threshold," i.e., an electrical stimulation threshold with light preceding each electrical stimulation, is determined. Next, another dark threshold is determined. This alternation of dark and light thresholds is continued throughout the session. The first threshold and the final threshold of every session are dark thresholds. Hence each light threshold may always be compared with the mean of two dark thresholds, i.e., those immediately preceding and following it. A one-minute rest period separates the succeeding threshold determinations, and no rest periods are allowed during any single determination. As previously described, intervals of 7.65

sec. separate stimuli in the dark condition and 15.3 sec. in the light.

The time required for a single session varies considerably, depending upon the number of light thresholds determined and the time required for each threshold. In general, the sessions were designed to last one to two hours, including the period of dark adaptation. Occasionally longer sessions were undertaken, in order to get a large number of points in one session. No single session exceeded two and one-half hours in duration.

Computation of the effects of light. The effect of a light flash on electrical sensitivity will be designated in the present experiments in terms of an enhancement value ϵ . Enhancement is defined as follows:

$$\epsilon = \log \frac{I_d}{I_l}$$

or, more simply,

$$\epsilon = \log I_d - \log I_l$$

where I_d is the mean of the two "dark" thresholds (i.e., the electrical thresholds determined immediately preceding and following the determination of the "light" threshold) and I_l is the "light" threshold itself. All values of I are in terms of electric current in microamperes. It should be stressed that ϵ is calculated with reference to the dark values that surround the particular light threshold for which the value of ϵ is obtained. In this way, the effects of changes in electrical threshold which might occur during the session as a result of fatigue, etc., are minimized in the calculation of ϵ .

Differences from Motokawa's procedure. An important modification of Motokawa's procedure which has been introduced here is the practice of repeating dark, or reference, thresholds before and after each light determination, and the use of the mean of these two as a basis for calculating the effects of the light. Motokawa begins

⁸ Previous experiments (Achelis and Merkulow, 1; Lewis, 12) have shown that the electrical threshold does not follow a course similar to the light threshold during dark adaptation. It is fairly stabilized at 20 minutes.

each session with a dark determination, and occasionally repeats this throughout the session, but apparently does not do so systematically. He uses the original value for calculating effects of light, and the repetitions serve only as a check on the stability of the S . If a later dark threshold differs from the original by more than 10 per cent, Motokawa discards the data for that session. In the present experiments, however, no data are discarded, and a mean of the two dark thresholds is taken regardless of the magnitude of the discrepancy between them.

Instead of Motokawa's ζ , as an expression of the enhancing effect of light, ϵ values are used in the present experiments. ϵ has the advantage of being expressed in logarithmic units as deemed appropriate in these experiments. It carries with it the property that equal percentages of enhancement or depression of electrical excitability are represented by equal positive or negative values of ϵ . In other words a depression of the threshold of six step units or an enhancement of six step units has the same absolute value on a log scale when expressed as ϵ , but different values when expressed as ζ . This modification of Motokawa's procedure is not in the same category as those discussed earlier, however, since it involves no change in experimental procedure, but merely a slightly different way of expressing the data. It may be noted that ζ and ϵ each have a value of zero when no enhancement is present, and that ζ and ϵ have similar positive and negative values for all small amounts of enhancement or depression.

Experiment II. Enhancement Following Flashes of Light From a Small Central Field

A conclusion of major importance in Motokawa's work is that the course of enhancement following a flash of light reaches a peak at a certain delay time. The delay time for this peak is called the crest time, and its value is said to depend chiefly on the *color* of the preceding flash of light. Experiment II was designed to test the dependence of crest time on color in as direct a fashion as possible.

Procedure. Since there was not enough time in any one experimental session to vary both color and delay time, a single color was used on any given day. Enhancement values were determined for

this color at the three delay times (1, 2, and 3 sec.) designated by Motokawa as the crest times for red, green, and blue lights respectively. The results were then evaluated simply by observing the degree to which the use of a red light, for example, resulted in a greater enhancement value at one second than at two or three seconds when these three enhancement values were found within a given experiment session. The design of the experiment was balanced in such a way as to minimize the factor of order of presentation within each session and from one session to another.

Stimulating conditions. The lights whose effects upon electrical sensitivity were measured in this experiment were presented by means of the optical system previously described. The $2^{\circ}8'$ field was used, and the duration of the light was two seconds. The intensities of the red, green, blue, and white light were equated at approximately 13 ft.-L. by the use of neutral density filters. The specific red, green, and blue colors were achieved by the use of the selective filters described in the apparatus section of this paper. The delay between the end of the two-second light flash and the electrical stimulation was set at 1, 2, or 3 sec.

Experimental session. During a single session in this experiment, the threshold following stimulation with one color was determined for each of the three delay times. The procedure used for a single session is described in the preceding section. The LC method was used to determine all thresholds. The initial electrical stimulation in each threshold determination was kept at the same general level for the light and dark determinations. Each descending series of electrical stimuli following flashes of light was started at the same intensity as the preceding dark. The starting intensities for

the dark determinations were decreased or increased by one step in a systematic fashion through the four determinations of a session. The Ss were unaware of this manipulation of the intensity with which the descending series began.

Experimental design. Four Ss participated in Experiment II, which consisted of eight sessions. Each of the three colors and white light were used in two sessions. The delay times were presented in the order 1, 2, 3 sec. for one of these sessions, and 3, 2, 1 sec. for the other session, for a given color. The colors were presented in an ABCDDCBA order through the eight sessions. Each of the four Ss began the series with a different color. The first delay time used in the session, 1 or 3 seconds, was alternated for the eight sessions. That is, in Session 1 the first delay time was 1 sec.; in Session 2, it was 3 sec.; in Session 3, 1 sec., etc. Two of the Ss started with a 1-sec. delay in Session 1 and two began with a 3-sec. delay.

Results

Effects of delay time on enhancement. The results of Experiment II are expressed as enhancement values (ϵ) computed from each of the thresholds following stimulation by light. The data of each session are presented separately. Figure 6 presents the two replications for each color for each of the four Ss. The abscissa of each of the 16 graphs presented in Fig. 6 is the delay time, i.e., the time in seconds between the end of the 2-sec. flash of light and the onset of the 0.1-sec. pulse of electrical stimulation. The ordinate of each is ϵ , the enhancement of the electrical sensitivity by light as defined above. The points designated by xs represent the enhancement values obtained during the first session with a given color and the os the second session.

Each enhancement value is based on a single threshold in the "light" condition as compared with two "dark" thresholds, i.e., those immediately preceding and following the "light" series.

It should be pointed out again that Motokawa reports that the maximum enhancement effect for blue light reliably occurs at 3 sec., for green and white light at 2 sec., and for red light at 1 sec. *The data in Fig. 6 show no such consistent "crest times."* It is also evident that the light does not consistently have the effect of enhancing the sensitivity to electrical stimulation. Instead, many of the points in Fig. 6 show "negative enhancement," and many are approximately zero. The S LAR is the only one who shows consistently "positive enhancement."

In general, for all four Ss, the enhancement resulting from stimulation by blue light is greater than that from the other two colors or white light, whereas red, green, and white have approximately equal effects within the individual subjects.

Variability. It is evident from Fig. 6 that there is a good deal of session-to-session variability in the enhancement values for a given color and delay time. For each color and delay time, the two enhancement values obtained on separate days were indicated by xs and os. Table 8 gives the mean discrepancy for all corresponding values of ϵ . Each of these means is for differences, both absolute and algebraic, together with the standard deviation of the algebraic values. Also given is the mean of the discrepancies between "dark" thresholds obtained before and after each light threshold. It can be seen from Table 8 that there is no consistent tendency for the enhancement value obtained on the second day to be higher (i.e., with a positive mean difference) or lower (with a negative mean dif-

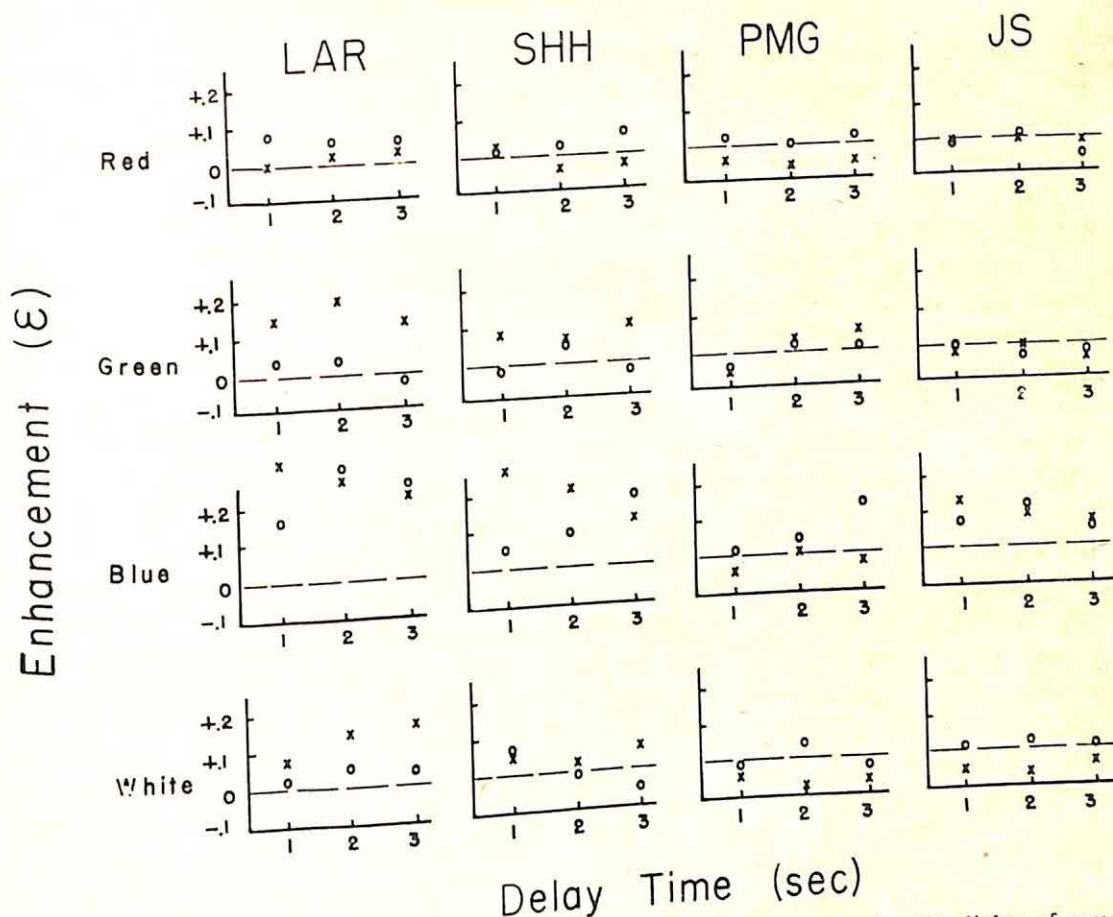


FIG. 6. Enhancement effects resulting from flashes of red, green, blue, and white lights of equal photometric luminance in Experiment II. The points designated by xs are for the first session and the os for the second session. The "delay times" of 1, 2, and 3 sec. are measured from the end of the 2 sec. flash of light to the beginning of the 0.1 sec. pulse of direct current. The light is presented as a $2^{\circ}8'$ field at approximately 13 ft.-L. Note that there is little agreement between these results and those of Motokawa as shown in Fig. 1. Where enhancement occurs at all it fails to conform to the specific wave-length curves that Motokawa has presented.

ference) than that of the first day.

The average discrepancy in the two

dark thresholds obtained before and after each light are also seen in Table 8. In

TABLE 8
VARIABILITY OF THE DATA OBTAINED IN EXPERIMENT II

(The differences in enhancement reported in this table are based on all 4 colors and 3 delay times, making a total of 12 differences going into each mean for each subject)

Subject	Differences Between Values Obtained Under the Same Conditions on Different Days			Mean Discrepancy (Absolute) Between Surrounding Dark Thresholds in Units of $2\frac{1}{2}\%$ Steps	Percentage of Dark Threshold Discrepancies of 5 Steps or Less
	Mean Difference (Absolute)	Mean Difference (Algebraic)	SD		
LAR	.087	+.053	.088	7.71	33%
SHH	.075	+.040	.090	5.17	58%
PMG	.050	-.045	.052	3.29	88%
JS	.027	-.008	.035	4.54	71%

the section of this report on evaluation of the LC method, similar measures of discrepancy were presented for Ss LAR and SHH. In that study, however, the thresholds compared were immediately adjacent, i.e., without an intervening light threshold. Under those conditions lower discrepancies were found, i.e., 4.77 steps in the earlier study as opposed to 7.71 steps just reported for LAR, and 3.41 as compared to 5.17 for SHH. The increased discrepancies probably reflect the effects of the longer time interval between dark determinations. For Ss PMG and JS, who did not take part in the earlier study, the mean discrepancy in the present study is less than five steps.

An important consideration in comparing these data to those of Motokawa is in the number of pairs of dark thresholds, preceding and following light, in which the discrepancy is less than 10 per cent, or approximately five steps. These figures also appear in Table 8. Motokawa discards data where the discrepancy between dark thresholds is greater than 10 per cent. It is evident that Motokawa would have rejected most of the data of S LAR, who showed the greatest enhancement values, but had only 8 out of 24 pairs of dark thresholds differing by five steps or less. On the other hand, 21 out of 24 of the enhancement points of PMG are based upon dark thresholds which vary by five steps or less, yet this S did not show consistently positive enhancement for any condition.

Conclusions with regard to enhancement as affected by delay time with a small central field. We may summarize the results of Experiment II, then, by stating that (a) positive values of enhancement were not universally obtained in this experiment, (b) there was a marked tendency for higher enhancement values to occur with flashes of blue

light than with white, green, or red when all flashes were of approximately equal photopic effectiveness, (c) only one subject (LAR) showed consistently positive enhancement values, and he was the most variable in his performance of all four of the Ss, and (d) there is no consistent tendency for red light to yield the greatest enhancement at a delay time of 1 sec., green light at 2 sec., or blue light at 3 sec.

Experiment III. Enhancement as A Function of Intensity of the Preceding Flash of Light from A Small Central Field

The preponderance of very small values of enhancement in Experiment II suggests that some of the specific conditions of the experiment may not have been optimal for enhancement effects to occur. This is in spite of the fact that the conditions appeared to approximate those used in the studies by Motokawa. Accordingly, it was decided to vary some aspects of the stimulus in order to test the possibility that other conditions might be more favorable. Motokawa has reported that the degree of enhancement is positively related to the intensity of the preceding flash of light. It was therefore decided to use the three Ss who showed relatively little enhancement in Experiment II, and find the degree of enhancement when intensity was varied. Red light and blue light were employed in Experiment III and three different intensities of each were presented. A 1-sec. delay time was used for the red and a 3-sec. time for the blue to favor a maximum degree of enhancement in accordance with the reports of Motokawa.

Procedure

Stimulating conditions. The conditions of Experiment III were similar to those of Experiment II. The flashes of

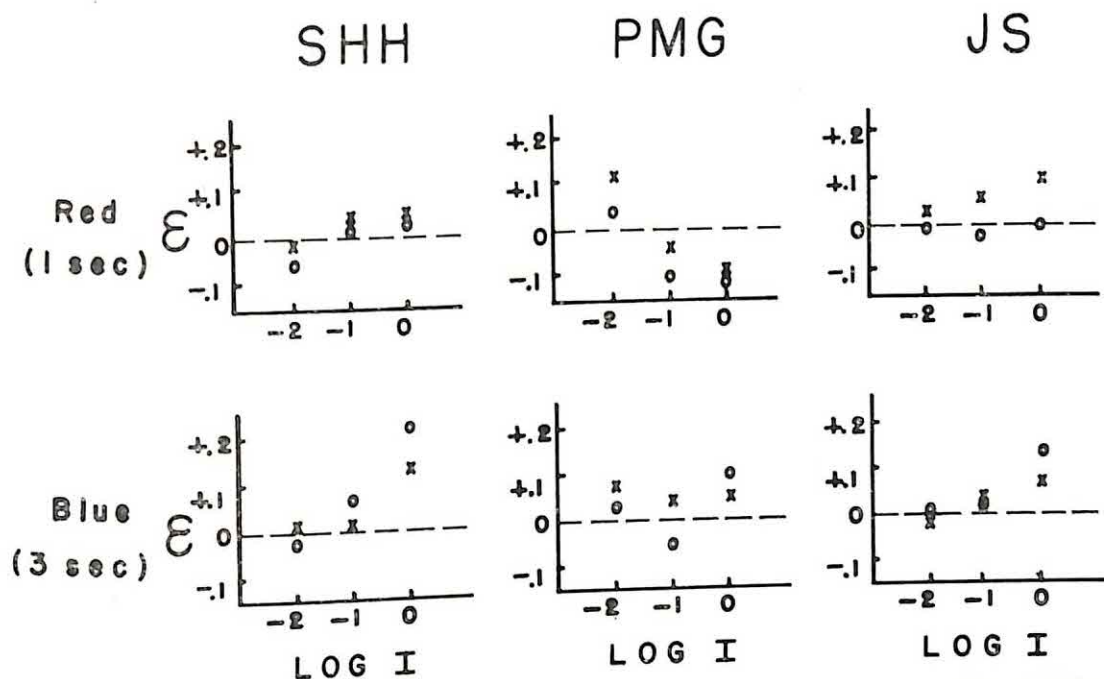


FIG. 7. Enhancement in relation to light intensity. Data for three different intensities of light, $2^{\circ}8'$ field, used in Experiment III.

light were seen as a $2^{\circ}8'$ field presented to the dark-adapted right eye of the S. Red (C) and blue (76) filters were used, and the variation in intensity of each was achieved by inserting 0.0, 1.0, or 2.0 log units of neutral density filter into the stimulating system.

During any one session, a single color was presented at each of the three intensities.

Experimental design. Three Ss served in this experiment for four sessions each. There were two sessions on each of the two colors, red and blue. During one of the sessions on each color, the intensities were presented in the order 0.0, 1.0, and 2.0, and during the other this order was reversed. The three Ss were those who showed least consistent enhancements in Experiment II.

Results. Figure 7 presents the data of Experiment III for each session for each S. The points marked with xs represent the first session on a given color, and os

the second session. It can be seen that S SHH shows an increased enhancement value with increasing intensity of light on all four sessions, with both blue and red light. The S JS has a clear intensity effect with blue light and on one session with red light. The S PMG, on the other hand, appears to have larger enhancement values for the lower intensity of red light, and the blue light has an irregular effect. Again in this study, the variability of the enhancement values under similar light conditions is high for all Ss. The variability of the dark thresholds preceding and following the light thresholds is of the same order as in Experiment II.

Experiment IV. Enhancement Produced By Matched Red and Blue Fields

Experiments II and III have yielded the information that blue light often produces considerably greater enhancement than does red of comparable photometric intensity. This finding suggests

that, despite the use of a small central field of stimulation, the effectiveness of these lights may be related to the extent to which they affect the rod, or scotopic response mechanisms of the retina. Experiment IV was set up to give further information about this point, and to provide further comparisons between delay times of 1 sec. and 3 sec.

Procedure

Stimulating conditions. In Experiment IV, two intensities of blue light and one of red light were used in each experimental session. The brighter blue light is a photopic match to the red, each flash having a luminance of about 13 ft.-L. The other blue light is less intense by about 2 log units; but its scotopic effectiveness is equivalent to that of the red light.

The light was presented as a $2^{\circ}8'$ field for two sec. The delay between the end of the light and the onset of electrical stimulation was either 1 sec. or 3 sec.; it was maintained at one or the other delay time throughout any given session.

Experimental design. Three Ss participated in four sessions each for this experiment. In each of the four sessions, the two intensities of blue light and the red light were presented. For two of the sessions, the delay time was 1 sec. and for the other two it was 3 sec. A single session began with one intensity of blue light, followed by red light, and ended with the other intensity of blue. The intensity of the initial blue light was varied in an ABBA order over the four sessions, while the delay times, 1 sec. and 3 sec., were alternated.

Results. Figure 8 presents the enhancement values obtained in each session. As before, the first session is indicated by xs and the second by os. The results for the 1- and 3-sec. delay times are pre-

sented separately. It is apparent that the less intense blue yields enhancement values more like those of the red light than does the more intense blue. In general, enhancements are greater for the more intense blue than for either the red or the dimmer blue. This suggests that the effectiveness of these lights may be more closely related to their scotopic effectiveness than to their luminance in foot-lamberts as given by a photopic function. For SHH and JS there is a clear increase in enhancement as intensity increases, supporting the earlier data on this point. For PMG, however, intensity does not have any consistent effect.

A comparison of the enhancements resulting from the 1 sec. and 3 sec. delay times does not yield the findings which would be predicted by Motokawa. That is, there is no consistent tendency for enhancement to be relatively greater at 3 sec. for blue light and at 1 sec. for red. The data appear to be slightly less variable with a 1-sec. delay than with a 3-sec. one.

The variability of the dark thresholds in this study is similar to that obtained earlier. For two Ss the average discrepancy between dark thresholds preceding and following the light threshold is slightly less than five steps, or 10 per cent, whereas this discrepancy is approximately seven steps for the third S, JS.

Experiment vs. Enhancement Following Flashes of Light from a Large Field

The results of Experiments II, III, and IV have shown that blue light usually produces a greater enhancement of electrical sensitivity than does green or red light of equal photometric luminance. This has led to the supposition that the enhancement due to a flash of light may be governed by its ability to stimulate the rod or scotopic receptor system. That

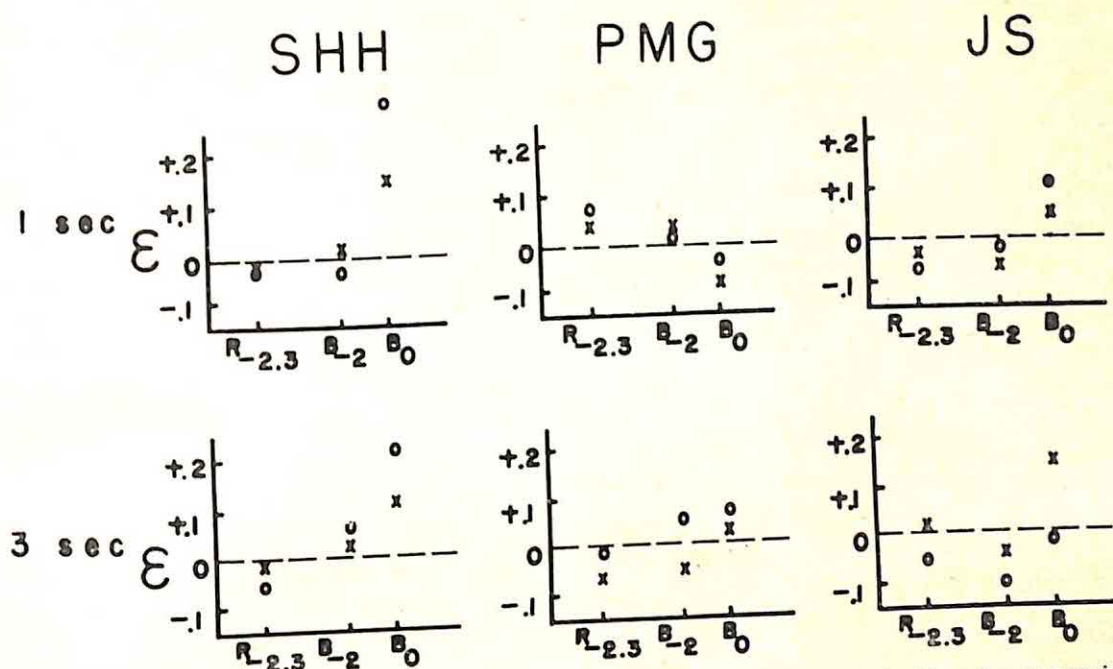


FIG. 8. Enhancement produced by matched red and blue fields. The red light, $R_{-2.3}$, has approximately the same luminance (13 ft.-L.) as the brighter blue light, B_0 . The scotopic effect of the red is approximately that of the dimmer blue light, B_{-2} . Results of Experiment IV with a $2^\circ 8'$ field and delay times of 1 sec. and 3 sec.

this should be true for stimulation by a small, central field of light may at first appear to be unlikely. Work with the human electroretinogram has clearly shown, however, that scotopic effects may arise in this way. Indeed it is found that it is difficult, if not impossible, to obtain any electrical responses from the stimulation by light of small central areas (Adrian, 2, 3; Riggs and Johnson, 24). Such responses as do occur in the human eye under these conditions have been shown by Boynton (4) to originate primarily from the action of stray light on peripheral scotopic receptors. These facts suggest that the peripheral regions of the retina occupy a preferred location, electrically, so that diffuse electrical stimulation of the eye by large electrodes has primarily peripheral effects.

If the enhancement resulting from small-field stimulation is due primarily to the action of stray light on peripheral

receptors, then it should be true that an even more effective means of securing enhancement would be to extend the size of field so as to provide a more direct stimulation of the rod receptors. Furthermore, most Ss report that the phosphenes aroused by weak electrical stimulation are not localized in the central so much as in the peripheral field of view. It seems appropriate, therefore, to use large-field stimulation in order to affect by light those regions of the retina that appear to be affected by electrical stimulation.⁹

⁹ We have found that when the intensity of the large field is reduced to near-threshold levels it becomes impossible for the S to distinguish between the "phosphenes" produced by photic and electrical stimuli, provided that the photic stimulation is in the form of a large field. Either form of stimulus results in a vague, colorless cloud of light.

Procedure

Stimulating conditions. For Experiment V, the light was presented as a 38° field, by means of the special eyepiece described in the apparatus section. Red, green, and blue lights, equated for scotopic effect, were used with central fixation.¹⁰

The delay time between the end of the light flash and the electrical stimulation was varied from .2 sec. to 4 sec. in order to cover the range of delay times reported by Motokawa (15) and by Mita, Hironaka, and Koike (13).

Experimental session. A single session in this experiment involved the presentation of a single color at four or five of the delay times. The duration of a single session was limited to $1\frac{1}{2}$ to $1\frac{3}{4}$ hours, and this factor determined the number of enhancement values that could be obtained. The delay times were divided into three categories. Category I included the short times (0.2, 0.4, 0.6, 0.8, and 1.0 sec.); category II included medium times (0.8, 1.0, 1.25, 1.50, and 2.0 sec.); and category III included times in integral seconds (1, 2, 3, and 4 sec.). On a given session the times from one of these categories were presented. Since the categories overlap, we have obtained several determinations for some of the delay times but only single determinations for others.

Experimental design. Four Ss participated in this experiment. Two of them, LAR and SHH, had already observed in the small field experiments. The other two had had no experience with the small field, but were given preliminary training in the observations of phos-

phenes with the 38° light field.

Each S required at least nine sessions to complete the experiment, since there were three colors and three time categories and each session involved a separate one of the nine possible combinations of time and color.

The order of presentation of the delay times within a single session was such that the shortest and longest delay times (within the category being presented) occurred near the middle of the session. Thus the session began and ended with intermediate rather than extreme values of delay time.

In general, all times categories for a single color were presented in consecutive experimental sessions. For example, one S started with three sessions in which red light was used, followed by three of green and three of blue. The order of presentation of colors and categories is included in Table 9 in the results section. The time category for the first session of a given color was randomized, as was the order of colors presented.

Results

Effects of delay time on enhancement. Table 9 shows the principal results of Experiment V. It is seen that the enhancement values are negative for short delay times for all Ss. Longer delay times yield predominantly positive enhancement for two of the Ss, but for the other two the enhancement values are essentially zero. In Fig. 9, an attempt is made to show these findings graphically. The number of experimental points is too great to be shown individually in this figure. Accordingly, all the ϵ values referring to the same condition of delay time, S, and color have been combined by averaging to a single value on the graph.¹¹

Our major conclusion from this ex-

¹⁰ Equal scotopic effectiveness was achieved for the C, G, and 76 filters, designated in this paper by R, G, and B, by the use of neutral density filters. As used here, the scotopic effectiveness is approximately equal to that of white at 3 ft.-L.

TABLE 9
ENHANCEMENT VALUES (ϵ) FOLLOWING FLASHES OF LIGHT FROM THE LARGE FIELD
USED IN EXPERIMENT V

Subject	Color	Session	Delay Time (sec.)									
			.2	.4	.6	.8	1.0	1.25	1.50	2.0	3.0	4.0
SHH	Green	4	-.251	+.017	+.076	+.140						
		5					+.117	+.153	+.077	+.165		
		6					+.126			+.165	+.095	+.126
	Red	3	-.326	+.111	+.155	+.311						
		1					-.007	+.218	+.178	+.176		
		2					+.220			+.148	+.160	+.118
MC	Blue	8	-.270	+.177	+.205	+.313						
		9					+.256	+.222	+.201	+.271		
		7					+.270			+.201	+.208	+.256
	Green	3	-.149	-.038	+.012	+.009	+.017					
		2				-.016	+.032	+.017	-.036	+.028		
		1					+.031			+.008	-.012	+.025
CJW	Red	5	-.342	-.075	-.057	-.086	-.018					
		9				-.058	-.009	-.062	-.046	-.044		
		4					-.014			-.013	-.030	+.010
	Blue	7	-.375	-.121	-.246	-.006	-.084					
		8				+.019	+.083	-.043	-.011	-.036		
		6					+.030			.000	-.032	-.040
LAR	Green	5	-.364	-.210	-.175	-.126	-.114					
		6				-.078	.000	-.070	-.021	+.017		
		7					-.086		-.021	-.029	-.077	-.116
	Red	10	-.274	-.253	-.131	-.135	-.112					
		9				-.078	-.056	-.065	-.057	-.081		
		8					-.074	+.003		+.128		
LAR	Blue	2					-.032			+.011	+.017	+.023
		3	-.211	-.170	-.078	-.046	-.057			-.043		
		4				-.004	-.045	-.022	-.017	+.048	+.032	+.042
	Green	10	-.061	+.009	+.106	+.240	+.182					
		11				+.143	+.125	+.152	+.172	+.225		
		3					+.148			+.010	+.185	+.024
LAR	Red	8	-.171	-.049	+.102	+.159	+.165					
		9				+.275	+.213	+.178	+.112	+.193		
		2					+.110			+.084	+.111	+.057
	Blue	5	-.157	+.169	+.242	+.304	+.206					
		7				+.341	+.247	+.194	+.309	+.225		
		6				+.269	+.096	+.184	+.124	+.236	+.188	+.262
LAR	Blue	1					+.304			+.209	+.242	+.255
		4					+.280					

¹¹The averaging has been done for convenience in presenting the results, and it must be realized that there is some variability in the points going into these averages. The individual points are found in Table 9. These points support equally well our conclusions with regard to enhancement as a function of delay time.

periment is that beyond about one second of delay time there is little relationship between delay time and degree of enhancement. This conclusion applies to each of the colors used; i.e., there is no apparent "crest time" as a function of

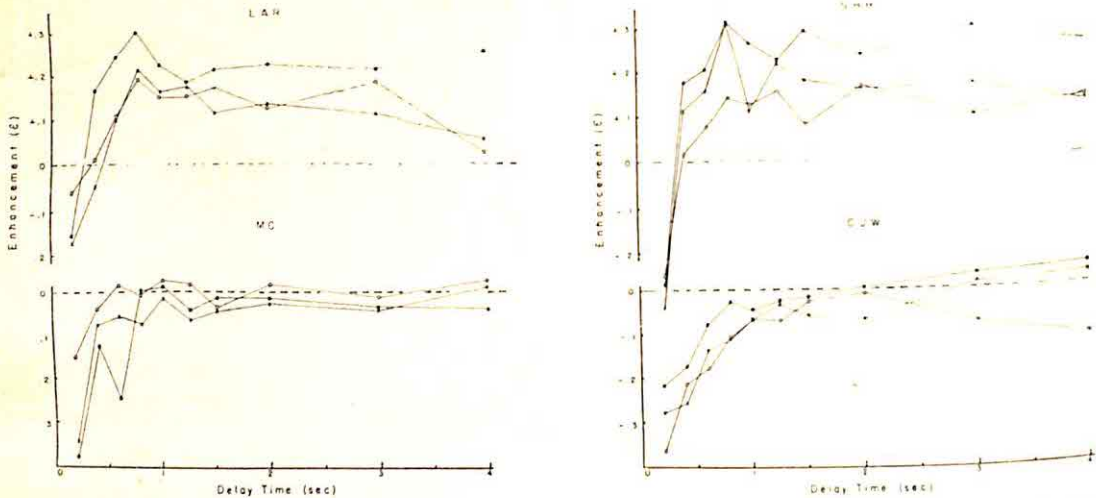


FIG. 9. Enhancement as a function of delay time for blue (\bullet), green (\circ), and red (\times) lights of equal scotopic effectiveness in Experiment V. The light is present as a 38° field whose scotopic effectiveness is similar to that of white light at 3 ft.-L. Compare with Motokawa's results in Figure 1. Note that enhancement is clearly present in but two out of the four subjects; and these two fail to conform to the pattern predicted by Motokawa. All four subjects, however, show clear evidence of reduced electrical sensitivity immediately following the light flash.

the wave length of the light. *There are no marked differences in effect among the three colors used here.* As before, however, the blue light does appear to have a greater effect than the red or green, especially in the cases of the two Ss, LAR and SHH, who showed large positive enhancements. This is in spite of the fact that the blue light is much less bright, in terms of photopic luminance, than the other colors. Since the three colors are approximately equal in their scotopic effectiveness, it appears to mean that the degree of enhancement is approximately what one might predict from a rod-receptor mechanism, but that blue is slightly more effective than would be expected.¹²

¹² This finding may perhaps be the result of Rayleigh scattering which, in the case of the human electroretinogram (Boynton, 4) has often given the appearance of excess blue sensitivity. Blue light is very effectively scattered by the optic media, so that it exerts a disproportionate influence on functions involving the extreme periphery of the eye.

Data on threshold variability. As in the previous studies, we have measured the number of steps separating the two reference thresholds used to compute the enhancement effect of each light. The number of such pairs for each S in Experiment V is from 36 to 51. For all Ss the average discrepancy is less than five steps, or approximately 10 per cent. Table 10 presents these averages, together with the percentages of pairs in which the discrepancy was five steps or less.

TABLE 10
AVERAGE DISCREPANCY BETWEEN REFERENCE
THRESHOLDS PRECEDING AND FOLLOWING A LIGHT THRESHOLD
(in 2 per cent step units)

	LAR	SHH	CJW	MC
Average Discrepancy	3.33	4.64	4.06	4.15
N	51	36	40	40
Percentage of Discrepancies of 5 Steps or Less.	79	64	70	75

The variability in dark thresholds in this experiment was in general smaller than had been found in the earlier studies. The reasons for this are not clear. Two of the Ss were highly experienced and two were relatively inexperienced. It is of interest that most of the pairs of reference thresholds (from 64 per cent to 79 per cent of them in the various Ss) had small enough discrepancies so that they would have been accepted as valid by Motokawa's criterion of reproducibility within 10 per cent.

Experiment VI. Enhancement as a Function of Intensity of Light from a Large Field

Experiments II, III, and IV demonstrated that with some Ss the degree of enhancement is positively related to intensity of light, as Motokawa has reported with a small, central field of light. Experiment V has demonstrated high values of enhancement for two of the four Ss when large-field stimulation is employed. Experiment VI is one in which a large field of white light is presented at three different levels of intensity. The aim of this experiment was to maximize the enhancement effect by utilizing a large field of light, and to study the influence of delay time and intensity on enhancement under these conditions.

Procedure

Stimulating conditions. All the data included in this section were obtained with the large (38°) field of light preceding the electrical pulse. Three intensities of white light were used. The middle intensity was the scotopic equivalent (with 2.6 log units of neutral density filter) of the colors presented in Experiment V. Its photometric luminance was approximately 3 ft.-L. The other two intensities were two log units above and

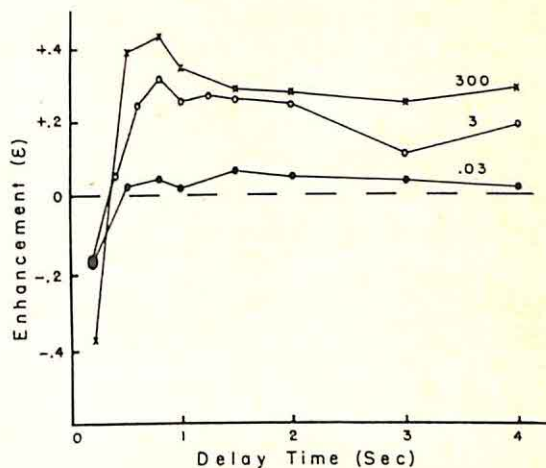


FIG. 10. Enhancement as a function of delay time for white lights of approximately 300, 3, and 0.03 ft.-L. luminance. Subject LAR, Experiment VI. Two other subjects failed to show consistently positive enhancement under the same conditions.

below that value. Each flash of light was followed, as before, by an electrical test stimulus at a delay time ranging from 0.2 sec. to 4 sec.

Experimental design. The design of Experiment VI was similar to that of Experiment V. In a given session, only one intensity of white light was used, at various delay times. Three Ss were used.

Results. As in previous experiments the thresholds for electrical stimulation were found to be high immediately following each flash of light. In other words, the value of ϵ was negative for delay times below about one sec. With longer delay times one S (LAR) showed positive values of ϵ as in earlier experiments. Two other Ss failed to show any positive enhancement effects.

Figure 10 presents the time course of enhancement for three intensities of white light for the one S whose electrical thresholds are markedly affected by the lights of various intensities. As in the previous experiment, some of the points on these curves represent the averages of

two or three determinations, while others are single values. It is clear that, for this S, the degree of enhancement is a function of the intensity of the light preceding the pulse. The most intense light yields enhancement values consistently greater than those for the other two intensities at all delay times except 0.2 sec., where the depressing effect of the intense light is greater than for the other two. The absolute values of these enhancements for the most intense light are higher than any previously obtained, and appear to reach a maximum in the neighborhood of one sec. after the flash. Some tendency toward the one-sec. maximum is also shown at the middle intensity of white light in this figure and in the most intense blue light of Fig. 9.

Experiment VII. Enhancement Produced by Large Fields of Colored Lights of Equal Luminance

Intensity of white light has been shown to determine the degree of enhancement in a S who readily shows enhancement effects. For colored lights it is clear from the data of Experiments II, III, IV, and V that this intensity should be evaluated not in photometric units of luminance but rather in terms of the effectiveness of the light for stimulating the scotopic system of the eye. Experiment VII is one in which this conclusion is further verified by the use of colored lights of approximately equal photopic luminance.

Experiment VII consisted of presenting blue, red, and green lights whose intensities were equivalent, photopically, to a white light of about 0.2 ft.-L (with 3.6 log units of neutral density filter). These lights were followed by electrical stimulation at 1-, 2-, 3-, and 4-sec. delay

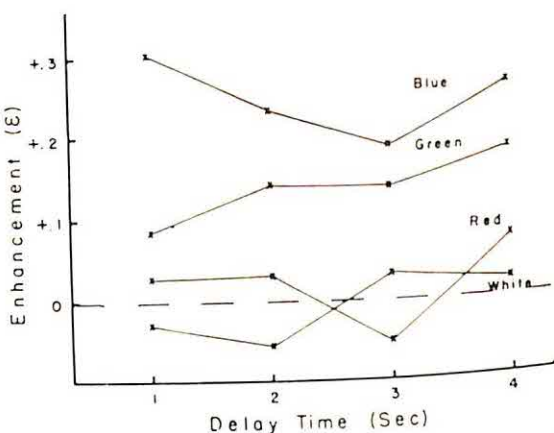


FIG. 11. Enhancement as a function of delay time for blue, green, red, and white lights of approximately 0.2 ft.-L. luminance. Subject LAR, Experiment VI. A second subject also showed greater enhancement for blue than for the other colors, but failed to show consistent effects in relation to the variable of delay time.

times. From the results of Experiments II, III, IV, and V it would be predicted that these lights, though of equal luminance, would yield different amounts of enhancement. Thus blue light, because of its relatively greater effectiveness for arousing the peripheral scotopic system, would be expected to produce greater enhancement than green, and green greater than red.

Figure 11 presents the results for the one S (LAR) who showed consistent enhancement effects. As predicted, the blue light is most effective in enhancing the electrical threshold, green light is less effective, and white and red lights are least effective. The points on these curves represent single determinations, and the three curves were obtained on consecutive days. The curve for white light was obtained approximately two weeks later. The data for another S were more variable but confirmed the fact that the blue light was the most effective of the colors used.

EFFECTS OF ELECTRICAL STIMULATION AT
SUBTHRESHOLD INTENSITY LEVELS*The Concept of Multiple Thresholds*

A central concept in the thinking and experimentation of Motokawa is that of multiple thresholds. This concept is to the effect that several electrical thresholds corresponding to the several fundamental response mechanisms of the retina may manifest themselves at once in the responses of the S. No provision for such a phenomenon is made in the usual psychophysical experiments. In fact, there appear to be certain logical difficulties implicit in the concept. Nevertheless, the stress laid on multiple thresholds by Motokawa and the revolutionary nature of this idea make it important that they be given serious consideration in relation to our procedure and results.

Specifically, *Motokawa assumes that several electrical thresholds may exist at any one time, one threshold for each of the fundamental response mechanisms affected by the preceding flash of light.* Of these multiple thresholds the true one is assumed to be the lowest, i.e., the one below which no phosphene is ever reported by the S. Motokawa (20) has insisted that great care is necessary to attain the lowermost or true threshold. In his words, "Trained Os can easily pass apparent thresholds and reach the true threshold, but untrained Os stop at one of the apparent thresholds, because they cannot discriminate sufficiently slight phosphenes from the background." An inexperienced experimenter may fail to continue presenting the stimuli in the descending series until the lowermost threshold is attained.

In order to avoid the danger of stopping the determination at one of the apparent thresholds

and not attaining the true one, Motokawa (21) has suggested the following procedure. When the threshold criterion has been reached, he presents further comparison pairs at an intensity level about 5 per cent lower than the threshold level. If the S can discriminate at this lower level, this is taken as an indication that the threshold attained was not the true one, but an apparent one, and the determination is continued until no lower discrimination is possible. More recently, Motokawa has devised another means of avoiding apparent thresholds. In a 1955 paper (22) he makes the following statement of his procedure: "In our routine work, we try to determine a true threshold alone, skipping over apparent ones by graduation of voltages in gross steps, but not forgetting that the obtained threshold is really a true one."

There appear to be certain logical difficulties in the concept of multiple thresholds. Let us assume that the blue response mechanism, for example, accounts for the lowermost or "true" electrical threshold and the red response mechanism has a somewhat higher electrical threshold under certain experimental conditions. Does this not mean that an electrical stimulus whose intensity is equal to that of the red response threshold will stimulate even more strongly the blue response mechanism? In other words, is it not difficult to conceive of a psychophysical relationship showing anything but a decreasing frequency of "yes" responses as stimulus intensity goes down, even though several different response mechanisms may be affected simultaneously?

In our own experiments the effects of light have been specified in terms of the change that the light produces in the threshold for electrical stimulation of the eye. Each threshold has been determined by the LC method as specified above in the section on general procedure. It must be emphasized that this method has been rigidly imposed in all cases, and that equal log steps have always been

TABLE 11

A. Total Number of Steps Between Original and Final Threshold

	Dark Condition					Light Condition				
	1	2	3	4	Total	1	2	3	4	Total
LAR	6	19	15	18	58	9	5	6	5	25
SHH	7	11	17	13	48	10	5	5	5	25
PMG	9	10	14	9	42	9	5	7	11	32
CJW	10	9	20	9	57	13	8	9	10	40
Total					205					122

B. Number of Steps Correctly Judged Between Original and Final Threshold

	Dark Condition					Light Condition				
	1	2	3	4	Total	1	2	3	4	Total
LAR	3	15	6	7	36	4	3	4	2	13
SHH	4	8	13	7	32	5	1	2	2	10
PMG	6	4	8	4	22	1	0	4	8	13
CJW	6	4	10	5	34	9	7	5	5	26
Total					124					62

extended threshold determinations was begun. The first and third of these were dark, or reference thresholds, and the second and fourth were light thresholds.

Results. The results of this experiment are of interest chiefly as comparisons of the dark and light conditions. Two such comparisons are presented. Both are based upon the responses made to stimuli *below* the level of the original threshold. That is, interest here is centered upon what happens during the special extension of the comparison stage of the LC method.

First, let us consider the number of two per cent steps between the original threshold and the termination of the extended threshold determination. It will be noted that the minimum number of steps between these two criteria is five, since the extended series was never terminated until at least five steps below the first threshold had been presented. Second, we may note the number of steps *correctly* judged between the original and final thresholds. No minimum number of

steps exists for this measure, since the *S* may make no correct judgments after his original threshold has been reached. This measure, then, reflects the success with which the *S* was able to discriminate at intensity levels below that which was called threshold in the earlier experiments. The data in Table 11 include both of these measures. Because the conditions (field size, delay time) differed slightly among the four *Ss*, individual data are presented in Table 11. All eight thresholds for any given *S* however were obtained under identical conditions. Both measures given in Table 11 show that *Ss* are often able to discriminate phosphenes from blanks at intensity levels below the original threshold in both the light and dark conditions. But in no case does the light condition exceed the dark condition in the number of successful discriminations. Instead, the dark condition seems to favor such discriminations, yielding about twice as many steps, when all *Ss* are combined, for both measures. This result is in contrast to the

prediction of Motokawa (22) for the multiple threshold effect. His prediction is that of a more abrupt transition from positive to negative responses in "dark" determinations as opposed to "light" determinations.

Experiment IX. Employment of the CS Method with Small Steps of Intensity Above and Below the Threshold

The results of Experiment VIII have indicated that phosphenes are sometimes aroused by stimuli below the usual threshold, but that this is not brought about in any unique fashion by the after-effects of flashes of light. The conventional concept of a threshold demands, in fact, that some positive responses be observed when stimuli of subthreshold intensity are given. Accordingly, Experiment IX was conducted to quantify the relationship between the frequency of positive responses and the intensity of stimulation over a wide range of subthreshold and suprathreshold intensities.

Procedure. The method of constant stimuli was used in Experiment IX. Nineteen intensities of electric current were presented 20 times each. These intensities differed by steps of two per cent, as was true of the steps used in the LC method. A blank stimulus was also presented 20 times, making a total of 400 presentations. A rest period of about one minute followed each 40 presentations. A random order of 300 stimuli, including each intensity 10 times, was presented twice in opposite directions. The S responded to each stimulus by saying "yes" or "no" and the experimenter made no comments except to identify the blank stimuli after they had occurred.

A 2-sec. blue light (76 plus 1.0 log units N. D.) preceded each electrical stimulation by a delay time of .5 sec. As in all of the experiments in which light

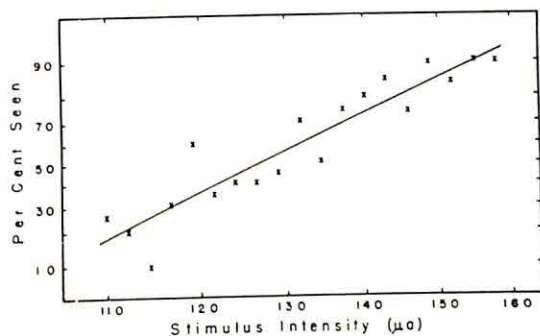


FIG. 12. Percentage of "yes" judgments for the appearance of phosphenes in relation to intensity of the stimulating current.

Method of constant stimuli with 2 per cent steps. Subject LAR, Experiment IX. This experiment was attempted with two additional subjects. Their data failed to satisfy the usual criterion of 75% or more judgments of "seen" at the high end and 25% or fewer at the low end. Presumably this is a consequence of the length of the experiment (2½ hours) and the small size of step (2%), factors that render the experiment an unusually arduous one for the subject.

was presented, 15.3 sec. separated successive stimulations. The small ($2^{\circ} 8'$) field of light was used, with central fixation.

The session was begun with a preliminary threshold determination by the LC method without light, followed by another preliminary LC threshold with the same intensity and delay time as was used in the CS determination. The CS determination was begun at approximately 30 minutes after the S entered the dark room, and it took approximately two hours to complete the series of stimulus presentations.

Results. Figure 12 presents the constant stimulus function for S LAR, plotted on probability paper. The abscissa is log intensity of current and the ordinate is per cent seen plotted on a scale of probability. The blank stimulus was never reported as seen during the 20 presentations. The straight line has been fitted to the points by the method of least squares.

Despite the small range of intensities used here (0.157 log unit), a wide range of response frequencies was obtained. Each of the three lowest steps yielded 25 per cent or fewer judgments of "yes" and the seven highest intensities were all "seen" on 75 per cent or more of the presentations.¹³ Although there are reversals in the function, there is no evidence of major dips and peaks, as would be predicted by the multiple threshold hypothesis. Rather, the points follow the prediction of a Gaussian function, with some variability which too would be predicted with such small intensity differences as 2 per cent between steps.

It is concluded, then, that the data of Experiments VIII and IX offer no support for the concept of multiple thresholds as outlined by Motokawa. Responses to subthreshold intensities of current are indeed observed, but the percentage of positive responses appears to exhibit an approximately Gaussian relationship to the log of the stimulating current.

SUMMARY OF RESULTS

A brief summary of the results of these experiments is as follows:

Experiments have been conducted in which the human eye was stimulated electrically by brief square-wave pulses. Three or four highly trained Ss were used in each experiment. The S judged the appearance or nonappearance of a phosphene in response to the pulse. A psychophysical method developed by Motokawa was used to determine the threshold for the electrical stimulation. This method involves successive stimulus

presentations in a descending series of intensities, as in the method of limits, followed by paired comparisons of blanks with successively weaker stimuli. When conducted under rigidly prescribed conditions this limits-comparison procedure can be used to measure changes in electrical sensitivity following flashes of light in the dark-adapted eye. More conventional psychophysical methods appear to be more reliable for determining electrical thresholds but are too time consuming to be used in extensive series of threshold dominations.

Immediately following a flash of light the eye is relatively inexcitable by electric current, but within less than one second after the light goes off the excitability rises to a normal or supranormal level. Supranormal excitability, or enhancement, is not found in some Ss but is clearly present in others. A study of various colors, areas, and intensities of stimulation has led us to conclude that the degree of enhancement depends largely on the extent to which the light has stimulated the rod, or scotopic receptor mechanism, of the eye. The enhancement effect often persists for several seconds after a flash of light. We have not found any clear peak, however, in the time course of such enhancement, and we have not found the specific wave-length effects reported by Motokawa for this time course. Certain factors of experimental procedure are believed to account for this discrepancy in results.

Discussion

The principal purpose of these experiments has been to evaluate the method of electrical stimulation of the eye as a tool for studying the mechanism of human color vision. After a general evaluation of methods of determining electrical thresholds (Experiment I), our primary

¹³ Attempts to obtain a constant stimulus function with 2 per cent steps, under similar light conditions, were made with two other Ss PMG and SHH. In both cases, however, the attempts were unsuccessful in that the range of intensities was not large enough to obtain an appropriate range of percentages seen.

concern was to attempt to verify the results reported by Motokawa for the enhancement of electrical excitability by flashes of lights of various wave lengths. Later we were led to a new series of experiments in which we explored the role of the scotopic or rod response mechanism in the enhancement of electrical sensitivity.

Comparison With the Motokawa Experiments

Experiments II, III, and IV were designed to follow closely the procedures reported by Motokawa. Red, green, blue, and white lights were used with central fixation on a small field with luminances within the range that Motokawa had reported as effective. Duration of the light flash was two seconds, the value most often used by him, and delay times of 1, 2, and 3 sec. were interposed between the flash and the 0.1-sec. square-wave test pulse of direct current through the eye. Three or four carefully trained Ss were used and adequate periods of dark adaptation and rest were provided.

Threshold determinations were made according to the design developed by Motokawa. A descending method of limits was followed by a comparing procedure in which the final threshold was ascertained by the S's failure to distinguish blanks from electrical pulses. Despite these similarities, we have not claimed that Motokawa's experiments have been exactly duplicated in our laboratory. The principal difference between our experiments and his appears to be in the application of the procedure in the daily routine of threshold determinations.

With such a high degree of similarity in experimental conditions we should certainly expect our own results to resemble those of Motokawa in all im-

portant respects. Nevertheless the extent of such agreement has been very limited, and may be described briefly as follows: First, we have indeed found some Ss who show impressive amounts of enhancement of electrical sensitivity resulting from flashes of light. The highest degree of enhancement seen under any of our conditions was one of about 0.4 log unit. This represents a drop in electrical threshold to approximately 40 per cent of its original value. In most cases, however, much smaller values of enhancement were obtained. In fact, nearly half of our Ss failed to show any consistent enhancement effects under any conditions. We have not found any explanation for these wide individual variations. Second, we have confirmed rather closely the early time course of enhancement, reported by Motokawa and his co-workers for the first half-second after the light goes off. Third, we have found, in some Ss, a clear relationship between the amount of enhancement and the intensity of the light flash that produced it.

While we have confirmed the existence of enhancement of electrical sensitivity by light, we have not been able to make use of this phenomenon as a tool for the study of color vision. This is because we have not obtained any differences among the various colors of light with regard to their aftereffects on electrical sensitivity of the eye. Motokawa's impressive series of studies, purporting to give an exact description of the fundamental response curves for human color vision, is based upon these specific aftereffects of various colors. Motokawa's consistency in obtaining these effects is in marked contrast to our own negative results, obtained under a variety of experimental conditions.

Differences in the routine of threshold determinations. We have corre-

sponded with Professor Motokawa about our negative findings with respect to specific wave-length effects. His opinion is that procedural differences are important in accounting for the differences in results obtained in the two laboratories. Motokawa has sent us sample protocols from one of his experiments, in order to illustrate the differences between his routine procedure and ours. With Motokawa's kind permission, we have reproduced one of these protocols as Table 12. We believe that it contains a more detailed description of his procedure than has appeared previously.

The protocol describes an experiment in which enhancement was measured following a flash of red light at various delay times. The typical time course was found, with the greatest enhancement occurring at a delay time of one second. The *S* was one of Motokawa's highly trained observers. We have taken the liberty of presenting Motokawa's protocol in such a way that the vertical position of each entry corresponds to the intensity of stimulation. The intensity values of the presentations are indicated on the left of the table, in units of electrical resistance in ohms. As the number of ohms increases, the stimulating voltage decreases.¹⁴ Spaced out in this way, the values clearly indicate the differences in step size which Motokawa has used in arriving at his thresholds. The columns headed $R_{0.5}$, R_1 , R_2 , and R_3 contain threshold determinations at delay times of 0.5, 1, 2, and 3 sec. after the cessation of a 2-sec. flash of red light. The columns headed R_0 indicate reference threshold determinations in which no light was present. The order of the columns indicates the order of threshold determinations in this experiment. Within each column are notations of the *Ss'* responses. An *o* in the body of the table is used to indicate that a phosphene was seen by the *S*, and an *x* indicates that it was not seen. The letters *w* and *vw* indicate the *S*'s report that the phosphene was weak or very weak. The notation *com* indicates the intensity at which the comparison stage of the procedure was demanded by the *S*. At all subsequent intensities, the comparison procedure was used. After the threshold was reached, several further comparison pairs

were presented, in order to ensure that the threshold obtained was the "true" one, and not an apparent one. This is indicated by a double *x*; and all further single *xs* indicate that the *S*'s response was "I don't know."

The concept of multiple thresholds, as outlined in a previous section of this paper, has led Motokawa to use gross steps in the initial phase of the determination of a threshold. Furthermore, in the determinations in which light was used, he employed gross steps in order to skip over the "dark" thresholds, i.e., the intensity values which had yielded thresholds with no light present. This procedure, also described by Motokawa in a recent publication (22), was instigated to avoid apparent thresholds and to ensure obtaining the true threshold.

In Experiments VIII and IX, some consequences of the multiple threshold concept were tested. The weight of the evidence appeared not to support this concept, but instead to be consistent with the usual view that a threshold is a point statistically defined within a transition zone of stimulus intensities over which the probability of a given response changes gradually from low to high.

The effect of step size on a threshold. Conventional treatments of psychophysics make no provision for multiple thresholds, assuming instead that a single threshold can be found by the use of equal steps along a continuum of stimulus intensities. In a descending series there is no exact point at which the *S* changes his judgment from "seen" to "not seen." Rather there is a transition zone from suprathreshold regions in which judgments of "seen" predominate, to subthreshold regions in which judgments of "not seen" become more and more numerous. The threshold is ordinarily defined statistically as that stim-

¹⁴ Motokawa's circuit for electrical stimulation can be found in reference 18.

TABLE 12

ADAPTATION OF A PROTOCOL BY MOTOKAWA (PRIVATE COMMUNICATION) SHOWING THE ROUTINE FOR DETERMINING ELECTRICAL THRESHOLDS. THRESHOLDS ARE INDICATED BY THE POSITION OF THE HORIZONTAL LINE NEAR THE BOTTOM OF EACH DATA COLUMN.

(See text for details)

	R_0	R_1	$R_{0.5}$	R_3	R_0	R_2
1000 -	o	o	o	o	o	o
2000 -	o	o	o	o	o	o
3000 -	o	o	o	o	o	o
4000 -	o	o	o	o	o	o
4500 -	o				o w	
5000 -	o com o <u>xxx</u> x	o w	o w	o	o o o o o com o o o o o <u>o o x</u>	o
5500 -	x x	o vw	o w o vw o com xxx x	o w o w o vw com o o o xxx x	xx x x	o o w o com
6000 -		o				xox o <u>xx</u> oxx x
6500 -		o o o com <u>xx</u> x				x
7000 -						

ulus value for which judgments of "seen" and "not seen" have an equal probability of occurrence.

The method of limits is one in which

the threshold is rather differently defined. As Guilford (9) has pointed out, much depends on the single terminal judgment of each series, at which the S's

judgment shifts from "seen" to "not seen" or vice versa. The limits-comparison method of Motokawa has the advantage that it removes some of this emphasis on single judgments by providing that each series end only after the S has tried unsuccessfully to make several comparative judgments between the stimulus and a blank. There is still a considerable transition zone, however, within which this criterion of unsuccessful responses can be met; and the probability that it will be met at any time during a descending series increases as the stimulus intensity is reduced. It is also obvious that the *number* of unsuccessful responses is determined, at any point within the transition zone, by the total number of stimuli presented at that point. The use of gross steps entails the presentation of fewer stimuli within a given region, while finer steps increase the number presented. Therefore, in a descending method such as this, the threshold criterion of unsuccessful responses is most likely to be satisfied in a region of the transition zone in which fine steps are used.¹⁵

It is our contention that, in the determination of electrical thresholds, the probability of finding a threshold at a given point is increased by the use of fine steps in that region of stimulus intensities. Motokawa has realized this in his practice of using gross steps for "skipping over" the unwanted dark and apparent thresholds in his search for the

true threshold. He has used fine steps systematically in the region where the true threshold is expected to be found. It is evident from the protocol in Table 12 that an elastic procedure has been used to select the size of step interval from one part of the session to another. In the limits stage of the determination, Motokawa's routine calls for presenting different steps of intensity for the "dark" series and the series with light. He has also varied these steps for the different delay times used within the "light" series. After comparing is begun, step intervals of 50, 100, or 200 ohms have been used arbitrarily from one time to another. We believe that the effect of this procedure is to heighten the probability that large values of enhancement will be shown for those particular series in which gross steps have been continued to relatively low intensity levels. A further effect is that of yielding the remarkable degree of consistency that the "dark" thresholds have shown in the Motokawa laboratory. We have never been able to achieve enhancement values conforming to such smooth functions as are shown in Motokawa's papers.

Further confirmation of the importance of step size is provided by an experiment conducted in Motokawa's laboratory. Through the kindness of Motokawa we have illustrated the results of one of his experiments in Table 13. In this special experiment the series of descending stimulus intensities has contained approximately equal log intervals. Again the first column lists the resistance values, in ohms, used to control the stimulus intensity. As in Table 12, ascending values of resistance correspond to descending values of stimulating voltage.

The results shown in Table 13 indicate that Motokawa, too, fails to find a specific crest time when equal log steps

¹⁵ Guilford (9) has discussed a somewhat analogous situation in the determination of a difference limen. He says, "Having a preconceived notion that Weber's law holds, E will use smaller steps in the lower part of the scale and larger steps in the upper part of the scale. It is conceivable that, with a highly experienced O or E, unless measures are taken to prevent it, Weber's law or any other law could be demonstrated to hold merely by a systematic choice of steps."

TABLE 13

ADAPTATION OF A PROTOCOL BY MOTOKAWA (PRIVATE COMMUNICATION) SHOWING A SPECIAL EXPERIMENT IN WHICH EQUAL LOG STEPS WERE EMPLOYED

(See text for details)

	R_0	R_0	R_1	R_0	R_2	R_0
2000 -	o w o w o vs o w o w o w	o o o o o o	o o o o o o	o o o w o w o w o w	o o o o w o o w	o o o o o o
2500 -	o w o w o com xox	o w o w o w o w o com	o o o o ?xo com	o o w o o o	o w o w o w o w o	o o o o o o
3000 -	o o o o o	o o o xoo o	o o o o o	o o com o xoo o	o o o com o o	o w o w o w o vw o com
3500 -	o o o o	o o o o	o ? xx x	o ? xx o	xoo xx x	o o o
4000 -	o xox x	o xx x	o o xx	xx oo xx	x x x	o oo xx
4500 -	x x	x x	x x x	x x x	x x x	x x x
5000 -			x x		x x	
5500 -			x		x	

of intensity are used. That is, thresholds obtained in the dark (R_0) condition are no higher than those following flashes of red light by one (R_1) or two (R_2) seconds. Consequently Motokawa concludes, and we agree, that no specific enhancement is revealed by red light with a one-second delay time under the conditions represented by Table 13. Motokawa has interpreted this to mean that the method of equal log steps involves too many presentations of electrical stimuli above

the threshold, with a resulting long-term insensitivity to the very weak stimuli that lead to such low thresholds as that of R_1 in Table 12. He has supported this interpretation by citing his most recent experiments (personal communication) which show that repeated high-intensity electrical stimuli produce long-lasting elevations in electrical threshold.

Our own interpretation is simply that the crest-time phenomena appear only when higher levels are skipped over and

stimuli are massed, by the use of small steps, at the lower levels. No long-lasting changes in electrical threshold have appeared in our own experiments. Such effects are negated by the following considerations: (a) Polarization effects or other changes in skin-electrode resistance are minimized by our use of the 200,000-ohm resistance in series with the S. (b) The data of Experiment I do not reveal a rising trend with repeated stimulation. (c) In our control experiment, thresholds obtained with a 7.65-sec. cycle were not higher than thresholds obtained with a 15.3-sec. cycle.

The Achromatic Nature of Our Enhancement Effects

Such enhancement effects as we have been able to observe appear to originate not from the color response mechanisms but from the scotopic system of the eye. This was first apparent in the course of Experiments II, III, and IV in which small central fields of light were used. It was under precisely these conditions that Motokawa obtained his most clear-cut results on the relationship of the time-course of enhancement to the wave length of the stimulating light. Experiments V, VI, and VII showed that even more marked enhancement effects occurred when the field of light was extended to cover a wide retinal area. In this case the scotopic nature of the effect

was even more readily apparent. It was concluded that in all of these cases the degree of enhancement was related to the extent to which the rod-receptor mechanisms of the eye were affected by the light.¹⁶ It was assumed that stray light acting on peripheral receptors accounted for these effects under the small central area condition of stimulation.

It is our conviction that the specific wave-length effects that Motokawa has described are the result of his very special method of experimenting in which the protocol is modified in accordance with the responses of the S. We have been unwilling to depart from a rigidly defined protocol in our threshold determinations, fearing that any such departure might influence the results in favor of some preconceived hypothesis. We are ready to admit that our negative results with regard to specific wave-length effects may simply mean that these effects are so subtle that they are masked by the ordinary variability of our threshold determinations. If this is so, however, we can hardly endorse the use of this method for studying the basic characteristics of color vision.

¹⁶ Consistent with this view is the fact that the phosphenes themselves appear colorless; this suggests that the electrical current does not affect the color response mechanisms when intensities of stimulation are low. It is likewise true that Ss typically report that the phosphenes appear in the periphery of the visual field, where relatively few color receptors are found.

REFERENCES

1. ACHLIS, J. D., & MERKULOW, J. Die elektrische Erregbarkeit des menschlichen Auges während der Dunkeladaptation. *Z. Sinnesphysiol.*, 1930, 60, 95-125.
2. ADRIAN, E. D. The electric response of the human eye. *J. Physiol.*, 1945, 104, 84-104.
3. ADRIAN, E. D. Rod and cone components in the electric response of the eye. *J. Physiol.*, 1946, 105, 24-37.
4. BOYNTON, R. M. Stray light and the human electroretinogram. *J. opt. Soc. Amer.*, 1953, 43, 442-449.
5. CLAUSEN, J. *Visual sensations (phosphenes) produced by AC sine wave stimulation*. Copenhagen: Munksgaard, 1955.
6. GEBHARD, J. W. Motokawa's studies on electric excitation of the human eye. *Psychol. Bull.*, 1953, 50, 73-111.
7. GRANIT, R. *Sensory mechanisms of the retina*. London: Oxford Univer. Press, 1947.
8. GRANIT, R. *Receptors and sensory perception*. New Haven: Yale Univer. Press, 1955.
9. GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.
10. HOWARTH, C. I. Strength duration curves for electrical stimulation of the human eye. *Quart. J. exp. Psychol.*, 1954, 6, 47-61.
11. JOHNSON, E. P. The electrical response of the human retina during dark adaptation. *J. exp. Psychol.*, 1949, 39, 597-609.
12. LEWIS, W. G. The effect of photic stimulation on the electrical phosphene threshold. Unpublished master's thesis, Brown Univer., 1953.
13. MITA, T., HIRONAKA, K., & KOIKE, I. The change in electrical excitability of the human retina caused by a flash of light. *Tohoku J. exp. Med.*, 1949, 51, 379-388.
14. MOTOKAWA, K. Visual function and the electrical excitability of the retina. *Tohoku J. exp. Med.*, 1949, 50, 307-318.
15. MOTOKAWA, K. Visual function and the electrical excitability of the retina. *Tohoku J. exp. Med.*, 1949, 51, 145-153.
16. MOTOKAWA, K. Electrophysiological studies of color vision. *Tohoku J. exp. Med.*, 1949, 51, 165-173.
17. MOTOKAWA, K. A physiological basis of color discrimination. *Tohoku J. exp. Med.*, 1949, 51, 197-205.
18. MOTOKAWA, K. Retinal processes and their role in color vision. *J. Neurophysiol.*, 1949, 12, 291-303.
19. MOTOKAWA, K. Physiological studies on mechanisms of color reception in normal and color-blind subjects. *J. Neurophysiol.*, 1949, 12, 465-474.
20. MOTOKAWA, K. Some remarks on measurements of electrical excitability of the human eye. *Tohoku J. exp. Med.*, 1951, 54, 385-392.
21. MOTOKAWA, K. Retinal traces and visual perception of movement. *J. exp. Psychol.*, 1953, 45, 369-377.
22. MOTOKAWA, K., & ISOBE, K. Spectral response curves and hue discrimination in normal and color-defective subjects. *J. opt. Soc. Amer.*, 1955, 45, 79-88.
23. RIGGS, L. A., BERRY, R. N., & WAYNER, M. A comparison of electrical and psychophysical determinations of the spectral sensitivity of the human eye. *J. opt. Soc. Amer.*, 1949, 39, 427-436.
24. RIGGS, L. A., & JOHNSON, E. P. Electrical responses of the human retina. *J. exp. Psychol.*, 1949, 39, 415-424.
25. WRIGHT, W. D. *Researches on normal and defective colour vision*. St. Louis: Mosby, 1947.

(Accepted for publication August 10, 1956)

A Factor-Analytic Study of Planning Abilities

R. M. BERGER,¹ J. P. GUILFORD, AND P. R. CHRISTENSEN¹*University of Southern California*

I. INTRODUCTION

THIS STUDY is the fifth in a series of investigations designed to explore abilities considered to be important in the successful performance of high-level personnel.² For the first four studies in

¹ Now at System Development Corporation (formerly System Development Division of the RAND Corporation).

² Under Contract N6onr-23810 with the Office of Naval Research. The ideas expressed in this paper are our own and do not necessarily represent the opinions of the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government. Of the authors of this paper, J. P. Guilford is responsible investigator and director of the project, P. R. Christensen is assistant director, and R. M. Berger was in direct charge of this particular study. For a fuller discussion of this study see references no. 2, 4, and 5.

P. R. Merrifield participated in the development of the ordering hypothesis and contributed greatly to the construction of items and the devising of scoring methods.

We are indebted to the following persons in the University of Southern California departments indicated for making available to us classes for preliminary experimental testing: Dr. Constance Lovell, Psychology; Dr. E. C. McDonagh, Sociology; Dr. D. W. Lefever and Mr. R. A. Jones, School of Education. In addition we gratefully acknowledge permission for test use and/or adaptation by the following authors and publishers: R. L. Thorndike (test 5); R. B. Cattell and the Psychological Corporation (test 6); L. G. Humphreys (tests 9, 12, 33, 34, 35); D. C. Adkins and Press Features, Inc. (test 21); D. C. Adkins and the AGO (test 22); and Irving Lorge (test 32). Tests 37, 39, 49, 50, and 52 were parts of the Aircrew Classification Battery.

We are very much indebted to the staff of the Personnel Research Laboratory, HRRC, Lackland AFB, Texas, for making the testing possible and for the expert administration of a long and demanding test battery. In particular we wish to

the series see (3, 6, 10, 13). The purpose of this study was to isolate and define the abilities involved in planning performances. We were especially interested in planning as it enters into the activities of supervisory, military, and research personnel. It is in the planning performances of these groups that we have looked for clues as to hypotheses of ability requirements that are common in almost any kind of planning regardless of vocational context.

There is little psychological literature bearing directly on planning; the most useful were the reports of previous factor analyses. A *planning* factor was found in the AAF Aviation Psychology Research Program but was not clear-cut as a very general or fundamental ability of planning or foresight (9). Analyses were made of two special AAF batteries of foresight-and-planning tests, both of which included a small number of planning tests plus reference tests. The planning tests helped to define two new factors: a *planning* factor and an *integration* factor. The agreement between the two analyses was not very strong with respect to the new factors.

Two alternative hypotheses were advanced in Report No. 5 of the *AAF Psychology Program Research Reports* (9) to explain the nature of the *planning* factor. One hypothesis was that it might represent a type of ideational fluency, i.e., with the idea that the man who can think of more solutions per unit of time would have an advantage with some of these tests. This explanation does not hold very well, since the production of ideas is no more important for performance on the tests that were loaded on the factor than for other tests that did not have significant loadings. The second hypothesis is

thank Dr. Lloyd G. Humphreys, Director of Research, Dr. J. W. Bowles, Technical Aide, and the military personnel who acted as test administrators and proctors.

that some form of visualization different from the manipulation type is required in tests of planning. Convincing evidence for this hypothesis is also lacking.

In an analysis of the AAF Sheppard Field battery of tests, Guilford, Fruchter, and Zimmerman (7) found a planning factor, which they labeled "planning speed." Most of the tests that appeared loaded on this factor had been analyzed in AAF studies discussed above. Due to the inclusion of a large number of planning tests, the investigators felt the *planning speed* factor was probably more stable than the *planning* factor found in the previous AAF analyses. The new factor was termed *planning speed*, since the emphasis had shifted to tests in which speed plays a more important role.

Lucas and French included several of the AAF planning and integration tests in a recent study (11). An integration factor corresponding to the AAF *integration III* was tentatively identified, although the planning tests did not show up on the factor as they did in AAF analyses. The presence of two planning tests on another factor suggested it might be a planning factor. The evidence is not clear, however, since the loadings and communalities tend to be very low.

II. HYPOTHESES

Planning was conceived not as a unitary psychological function but as a term to cover certain classes of activities. On the basis of our considerations of planning activities and previous studies, six major hypotheses were developed concerning the abilities involved. Four were formulated in terms of classes of activities involved in the various stages of planning. These were: *orientation*, hypothesized as an ability to see an order or trend in a mass of information and to recognize the pertinent variables operating in a situation; *prediction*, the ability to see beyond the material given; *elaboration*, the ability to produce ideas or representations of ideas that contribute to the development of plans; and *ordering*, the ability to arrange bits of information into a meaningful series. Two additional hypotheses were constructed pertaining to the qualitative aspects of planning. They were: *ingenuity*, the

ability to find new and different methods of accomplishing a task; and *evaluation*, the ability to detect limitations in plans and to determine relative importance of variables.

Each major hypothesis indicates an expected primary thinking ability. Each subhypothesis reflects a specific feature or variation or alternative conception of the primary-ability dimension. These hypotheses and subhypotheses have served as the logical base for what was recognized as a continuing research program in which this study is an exploratory investigation. From an operational point of view, the subhypotheses served as starting points for test ideas.

The complete list of hypotheses, subhypotheses, and the tests (by number) designed to measure them follows:

1. Orientation
 - a. Sensitivity to order: seeing an order or trend in a relatively ambiguous situation (tests 1-3, see Section III).
 - b. Recognition of variables: awareness of the pertinent variables operating in a situation (tests 4-5).
2. Prediction
 - a. Extrapolation: seeing the effects of a spatial or temporal extension of a present trend (tests 6-8).
 - b. Foresight: relating what is given to what lies ahead (tests 9-12).
3. Elaboration
 - a. Specification: specifying the necessary details (tests 13-15).
 - b. Production of alternate methods: finding different arrangements of the same material that will fulfill the requirements of an adequate solution (tests 16-17).
 - c. Symbolization: producing adequate representations of ideas (tests 18-19).
4. Ordering
 - a. Temporal ordering: arranging steps in a time sequence (tests 20-22).
 - b. Hierarchical ordering: arranging subject matter according to class and subclass (tests 23-25).
5. Ingenuity
 - a. New methods: finding new or uncommon procedures (tests 26-27).
 - b. New applications: adapting methods to new situations (test 28).

6. Evaluation

- a. Importance of variables: determining relative importance of variables (tests 29-30).
- b. Seeing deficiencies: detecting the shortcomings of a proposed procedure (test 31).

Reference factors (previously known factors expected) are: *planning* (AAF analyses), *integration III*, *verbal comprehension*, *general reasoning*, *logical evaluation*, *originality*, *ideational fluency*, *associational fluency*, *adaptive flexibility*, *eduction of conceptual relations*, *judgment*, *visualization*, *symbol manipulation*, and *numerical facility*.

Included in the battery is a test that is necessarily complex in that it requires the examinees to devise a complete plan. The test is an adaptation of Lorge's Planning Skills test. A detailed account is given of a military situation in which there are problems of morale and leadership; the examinee is required to write down the steps in a plan of his own making that will improve the situation. Further description of the test will be given in section III. It should be informative to note to what extent a score on a test of general planning will load on the various factors that emerge in the analysis.

III. THE TEST BATTERY

Several considerations influenced the kinds of new tests constructed and kinds of scores used:

1. Many of the tests require completion-type answers. It was felt that since some planning activities are necessarily productive, the measures of the abilities associated would best be made on the basis of productive efforts of the examinee.
2. An attempt was made to develop tests which, although of the completion type, might be scored objectively. A minimum requirement for those to be scored subjectively was that it was possible to develop a reasonably reliable score.
3. Separately timed parts were constructed where possible to permit the computation of meaningful alternate-forms reliability estimates.
4. An effort was made to construct unique

tests where possible, that is, tests of one common factor, except in a few cases where the task presented is of necessity complex.

5. Where possible, simple tasks and situations were used in order to minimize variances due to differences in personal experience.

Most of the new or adapted tests that were to serve as experimental tests were pretested on groups of college students in order to determine adequate lengths, time limits, and instructions. Information derived from this pretesting was used in rejecting some of the tests and in revising and improving the 31 experimental tests to be included in the final battery along with 16 reference tests. In addition, several Air Force test scores from the classification battery were included in the analysis, also for reference.

All tests included in the analysis are described in the next section. For convenience, they are arranged in the order of the hypotheses rather than in the order of administration. For each test, the nature of the task is stated briefly, and wherever possible a sample item is given.

It should be noted that the sample items are transparently easy, since they were used as part of the test instructions for orientation. With respect to difficulty, then, the sample items are not typical of the items in the test proper.

The number of separately timed parts, the number of items per part, and the working time are indicated. Since the scoring procedure for many of the tests, especially those requiring subjective evaluations of completion responses, may be of interest, a summary of the scoring protocol is also given for most of such tests.

The code numbers of the experimental tests and some reference tests developed at this laboratory indicate the following information: The first letter—P, R, C, or E—indicates planning, reasoning, creativity, or evaluation, respectively. The second letter stands for a specific hypothe-

sis; in the case of planning, R means orientation, X means prediction, E means elaboration, O means ordering, I means ingenuity, and J means evaluation. Next comes the number of the test under the hypothesis, usually in order of the sub-hypothesis listed: 01, 02, etc. The last letter indicates the form of the test, A meaning the first form, B the second form, etc. Several experimental tests are uncoded; these are tests adopted or only slightly adapted from other sources.

A. DESCRIPTION OF EXPERIMENTAL TESTS

1. Orientation

a. Sensitivity to Order

Test 1. Matrix Order—PRO1A. Each

A. mouse	rat	lion	pig	cow	horse	elephant
Answer: <i>Animals become larger.</i>						
B. century	year	decade	day	week	second	minute
Answer: <i>Time unit becomes smaller.</i>						

item is a 3×3 matrix of words. *E* (the examinee) is required to draw a straight line through the sequence of three words that are in the most meaningful order. The line may be drawn through any one of three columns, any one of three rows, or either of two diagonals. In addition, the direction of the order is indicated by arrowheads. An arrowhead at one end shows the temporal order of the sequence and arrowheads at both ends of the line signify a spatial or logical relationship that can go either way (see Fig. 1).

Sample items:

A. run	think	want
read	write	sleep
walk	publish	book

↓

B. coat	tank	cloth
accident	shirt	flare
mixmaster	valley	skin

↖ ↗

FIG. 1

An order in time is found in sample A. Order in space is found in the diagonal of sample B. The double arrow in sample B indicates that the sequence "skin-shirt-

coat" is as logical an order as "coat-shirt-skin."

Scoring: An item is counted correct only when both the correct line is drawn and the correct placement of the arrowhead is made. The total score is the number of items correctly marked.

Parts, 2; items per part, 15; working time, 18 minutes; maximum score (highest possible score), 30.

Test 2. Seeing Trends—RIO6A. Each item consists of a group of words whose arrangement is not entirely uniform. *E* attempts to discover a trend from left to right and writes his description of the trend below the item.

Sample items:

Note that in both items several reversals of the trend occur between words taken as pairs and it is necessary to look beyond adjacent words for the predominant trend.

Scoring: Total score is the number of correctly marked items. Only one kind of response is keyed for each item. In some cases responses worded differently than the keyed ones are acceptable if the meanings are close or if the responses are negatively worded.

Parts, 2; items per part, 10; working time, 10 minutes; maximum score, 20.

Test 3. Sensitivity to Order—PRO2A. Each item consists of 5 words that are ordered in some way. Some of the items present the words in the right order, others have one word out of order. If the arrangement as given is correct, *E* marks a plus sign. If the words are not in their best order *E* is required to circle the word that is out of place and to put a

check to show where that word belongs in the group (see Fig. 2).

Sample items:

A. key,	lock,	door,	room,	house	+
B. accident,	writeup,	reporter,	newspaper,	reader	—

FIG. 2

In sample item *A* the objects are in good order. In sample item *B* the arrangement is not as good as it might be, since the presence of a reporter is usually necessary before the write-up of an accident can be made. "Write-up" is circled and a check indicates that it belongs between reporter and newspaper.

Scoring: Only the items in which some change is to be indicated in the order are scored. The presence of correctly arranged items that the examinee is required to mark with a "plus" serves to keep the examinee from indicating changes in other items that also appear to be properly arranged to him. The items that are scored must have the appropriate word circled and the check correctly placed in order to receive credit. The total score is the number right.

Parts, 2; items per part, 15; working time, 8 minutes; maximum score, 18.

b. Recognition of Variables

Test 4. Pertinent Questions—PRO₃A. *E* is presented with a number of conflict situations in which additional information would be required for making a choice. *E* is asked to write four questions, the answers to which would serve as a basis for making a decision.

Sample item:

A student who has just graduated from college is offered positions in two different parts of the country. What four questions have to be considered in making a choice?

Possible answers: What would the work involve in the two jobs? What salaries would be paid? In what places are the jobs? What are the relative opportunities for the future?

Other questions may be acceptable but they may not repeat previous questions and they must refer to different aspects of the problem.

Scoring: One point is given for each response that refers to a separate aspect of the problem. The total score is the number of responses credited.

Parts, 2; items per part, 4; working time, 12 minutes; maximum score, 32.

Test 5. Awareness of Variables—PRO₄A. This is a test developed by the Human Resources Research Laboratories at Bolling AFB. A problem is presented in which some decision is needed. Accompanying the problem are a number of statements of facts that may be pertinent to the problem. *E*'s task is to select a specified number of facts that seem most important in reaching a decision.

Sample item:

You are driving from Middleton to Warrensburg. You have plenty of time for the trip. There are two possible routes: Route A and Route B. Which would be better to take?

- Route A is 105 miles long and Route B is 115 miles long.
- Route A goes through 5 small cities on the way, while Route B goes through 3 somewhat larger cities.
- Route A has 10 miles of road under construction, with dirt surface and sections of one-way traffic.
- Route B is mostly winding, bumpy macadam, while Route A is mostly 4-strip concrete.
- Route A goes north of Lake Winepaw, while Route B goes south of the lake.

Which two facts are the most important to consider in deciding which route to take? Circle the letters preceding the two statements selected.

The best answers are generally agreed to be *C* and *D*.

In the test itself, each problem has 15 facts under it, and *E* is to select the five facts that are most important.

Scoring: Expert opinion was used in determining the "most important" facts for each of the four problems; the facts

on which most experts agreed constituted the provisional key. Scores were obtained using this key and the results were subjected to item analysis. This led to certain minor modifications resulting in the final key. The number of "important" facts in this final key varies from 4 to 5, depending on the empirical results. In scoring, a weight of 1 is given to each agreement with the key. The total score is the number of such agreements.

Parts, 4; items per part, 1; elements per item, 15; working time, 12 minutes; maximum score, 18.

2. Prediction

a. Extrapolation

Test 6. Series. This test is an adaptation from the items included in a part of R. B. Cattell's *A Culture-Free Test*. Each item consists of a series of figures, which are varied in some systematic fashion. *E* has to select the figure that goes next in the series by choosing among six given alternatives (see Fig. 3).

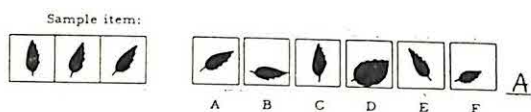


FIG. 3

Scoring: Score is the number of items correctly marked.

Parts, 2; items per part, Part 1-7, Part 2-8; working time, 5½ minutes; maximum score, 15.

Test 7. Effects-PXOrA. Given a statement of some present-day trend, *E* attempts to predict what the nature of the effects of this trend will be at a specified future time. Each item consists of the statement of a trend and provides room for writing down four possible effects. The trends given are realistic and cover many different areas in which there is common experience.

Sample item:

A. There have been more girls born in the last five years than boys. What effects will this have 20 years from now (in addition to the obvious effect that there will be more women than men)?

Possible answers: *More unmarried women than unmarried men. More rivalry among women for husbands. More women would take up career work. Agitation to allow polygamous marriage.*

Scoring: Four responses can be given to each item and each response is weighted either 0, 1, or 2. If the response is a duplication of a previously given one or is not a possible effect of the trend given, it is weighted zero. Effects that are possible are weighted 1; an additional weight of 1, or a total response weight of 2, is allowed if the effect is less obvious or is somewhat indirect. The total score is the sum of the weights.

Parts, 2; items per part, 5; working time, 16 minutes; maximum score, 80.

Test 8. Consequences (Remoteness)-CFO6A. For each item in this test a change of an unusual nature is imposed which would disrupt the normal state of affairs. *E* is required to list as many different consequences or results of these changes as he can.

Sample item:

What would happen if all national and local laws were suddenly abolished?

Possible answers: *People would no longer get parking tickets. There would be more killing. No more taxes.*

Scoring: There are two scores derived from this test: a low-quality score and a remoteness score. For the low-quality score the number of direct responses is determined. Direct responses are those that describe direct or immediate consequences of the given changes. The low-quality score is used as a measure of *ideational fluency* (see test 44). The remoteness score is the one used here as a meas-

ure of *prediction*. The score is based on the number of indirect or remote consequences listed. The individual responses are rated as direct or remote by comparing them with a comprehensive list of model consequences that had been shown to be reliably weighted in a previous scoring for the creativity analysis. Every response is thus scored either 1 or 0 for remoteness. The total score is the sum of the weights.

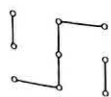
Parts, 4; items per part, 1; working time, 8 minutes; maximum score, unrestricted (U).

b. Foresight

Test 9. Competitive Planning. This is an Air Force test modified slightly for purposes of this study. The test is based on the familiar "completion-of-squares" game. Each of two hypothetical opponents (Black and White) tries to complete as many squares as possible in figures that are incomplete. *E* has to plan the moves for both opponents so that each achieves the maximum number of squares possible. Black always makes the first move, filling in one side of an incomplete square. Each time Black or White completes a square, he has to make one additional move. *E* must look ahead in many cases in order to see that a move giving an immediate gain should be passed over in favor of one making possible greater gains for future moves.

This is an answer-sheet test and an answer legend provides five alternative outcomes for the number of squares completed by the opponents, Black and White (see Fig. 4).

Sample item:



Answer Legend

A.	Black 0	White 4
B.	Black 1	White 3
C.	Black 2	White 2
D.	Black 3	White 1
E.	Black 4	White 0

FIG. 4

Making the best moves possible for their individual interests, Black and White complete this sample problem in the following manner: as always, the first move is made by Black. No matter which side Black fills in, White is able to complete two squares immediately. After completing the second square, White fills in a side of one of the remaining squares. This enables Black to complete the remaining squares so that the final result becomes two squares for Black and two squares for White. The answer marked, therefore, should be C.

Scoring: The score is the number of items right minus the number wrong divided by four ($R - \frac{W}{4}$).

Parts, 2; items per part, 20; working time, 20 minutes; maximum score, 40.

Test 10. Symbol Grouping-PXO2A. Each item consists of a line of X's, -'s, and O's presented in a scrambled order. The task is to make successive moves of groups of one or more adjacent symbols to achieve the correct order. The desired order for all items is first all the X's, next all the -'s, and last all the O's. The score depends not only on finding a solution but also on making as few moves as possible. Since there are a variety of ways of getting to the correct order, foresight is hypothesized to be demonstrated when the most economical procedure is adopted. *E* indicates each move by circling the symbol or symbols to be moved and placing a checkmark at the place they are to go. This is repeated for each new arrangement resulting until the solution is reached (see Fig. 5).

Scoring: A weight of 2 is given when the solution is reached in the minimum number of steps, and a weight of 1 is given for the minimum number of steps plus one more step. Two steps or more

Sample item:

Given: X - O X X - O X -, the following moves lead to the correct order.

x✓ - o x x - o (x -)
 x x✓ - o (x x) - o
 x x x x - - - o o

FIG. 5

beyond the minimum is felt to indicate little or no foresight and is therefore given no credit. Unsolved and unattempted problems are also not credited. The total score is the sum of the weighted item scores.

Parts, 2; items per part, 8; working time, 10 minutes; maximum score, 32.

Test 11. Contingencies-PXO₃A. In each item an object is named that is to be considered in connection with a given situation. *E*'s task is to state a possible condition that might require the use of a given object in that situation. Four objects are presented with each item in the test.

Sample item:

Sally and Jane are going out to a farm to pick berries for the day. Their pay will be half the berries they pick. Included in the things they will take are the objects listed below. For each object, state a condition or circumstance that may arise requiring the use of that object. (Do not merely give the use of the object, but write why it might come into use.)

Ointment If bitten by insects or scratched.
 Pins For possible rips in the clothes.

Scoring: Each written response is scored as either acceptable or unacceptable. A response is acceptable if a plausible condition is given that might arise and require the use of the object. The total score is the number of acceptable responses.

Parts, 2; items per part, 20; working time, 12 minutes; maximum score, 40.

Test 12. Route Planning-CI₄IIAX. This is an adaptation of an Air Force test. For each item *E* must trace a path from the corner of a printed maze to a

goal box in its center. There are four item numbers, one at each corner of the maze, that serve as starting points. The task is to locate the one point through which one must pass in going to the goal. Letters mark the points on the various routes between the starting point and the goal. When *E* finds the one letter which he must pass through in order to reach the goal he marks the letter on the special answer sheet provided (see Fig. 6).

Sample item:

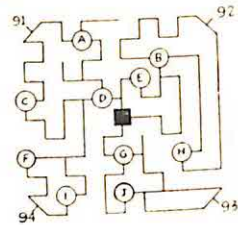


FIG. 6

In going from 91 to the goal, you may pass through either A or C; however, you must pass through D; therefore, D is the right answer. Answers for other items are: 92-B, 93-G, and 94-D.

Scoring: $R - W/2$.

Parts, 5; items per part, 8; working time, 17 minutes; maximum scores, 40.

3. Elaboration

a. Specification

Test 13. Planning Skills II-PEO₂A. This test is an adaptation of Irving Lorge's Planning Skills Test (the latter is employed as a test of general planning and described as test 32). *E* is allowed five minutes to read a description of the conditions prevailing at some hypothetical military, weather-station outpost in the Arctic. The conditions for the most part entail the effect of the rigors of Arctic temperature and isolation on personnel assigned to the stations. After reading the several paragraphs involved, *E* is given the signal to turn to the first part of the test, which requires that he write

as detailed a plan as possible on some given aspect of the problem situation. For the second half of the test another particular aspect of the problem situation is given and *E* is to write down all the details necessary for coping with that part of the problem.

Scoring: Each detail listed, if specific enough, is given credit. The total score is the number of acceptable details.

Parts, 2; items per part, 1; working time, 16 minutes; maximum score, U.

Test 14. Planning Elaboration—PEO₁₄A. Each part consists of a statement in which some plan of activity is briefly outlined. *E*'s task is to fill in as many of the details as he feels necessary to make the plan work. Each point or detail specified is written on a separate line.

Scoring: The total score is the number of details specified.

Parts, 2; items per part, 1; working time, 16 minutes; maximum score, U.

Test 15. Figure Production—PEO₁₅A. Each item in this test consists of markings of one type or another. The markings as given have no particular meaning. *E*'s task is to add to what is given as much as he feels necessary to produce a complete and meaningful figure. *E* can draw any figure he wishes as long as the given markings are definitely incorporated in what he makes (see Fig. 7).



FIG. 7

Scoring: A total of 140 sample responses to the stimuli were categorized by three raters into a 5-point equal-interval scale of "elaborateness." The responses on which all raters agreed were used as models for their respective categories. Categories were assigned weights from 0 (least elaboration) to 4 (greatest

elaboration). The remaining responses were weighted according to the category in which they were placed by a qualified rater. Score is the sum of the weights over all responses.

Parts, 3; items per part, 4; working time, 9 minutes; maximum score, 48.

b. Production of Alternate Methods

Test 16. Alternate Methods—PEO₁₆A. In each part a task is described that may be carried out in a number of nearly equally effective ways. *E* attempts to write down as many as six different ways of distributing the work and assigning personnel in order to accomplish the task.

Sample item:

A house located near a stream is on fire. Twenty men, each carrying a bucket, arrive to help put out the fire. The house is about 20 yards from the stream. In how many ways could you organize this bucket brigade to deal with the fire?

Possible answers: a. Form a line of 20 men; the first man fills a bucket and it is passed from man to man; the last man throws the water on the fire and runs back with the empty bucket.

b. Form two lines of 10 men. The filled buckets are passed up one line and the empty ones back down the other.

c. Assign 5 men to fill the buckets, 10 men to run with the buckets, and 5 men to throw the water where it is needed.

Scoring: In order to be counted as an acceptable response, each method listed must be practicable, employ all or most of the given material, and be in some way different from the other methods produced. The total score for the four items is the number of acceptable responses.

Parts, 4; items per part, 1; working time, 14 minutes; maximum score, 24.

Test 17. Match Problems II—PEO₁₇A. This test is a varied form of the Match Problems test to be described in a later section (test 46). These tests utilize the match-stick problems in which groups of matches are laid out to form patterns of squares or triangles. *E* must find a way

of taking away a specified number of matches to leave a certain number of squares or triangles remaining. In test 17 *E* has to find several different solutions for each problem. That is, different *patterns* of matches must be removed to leave the specified number of squares or triangles. Three or four different ways of solving each problem are possible (see Fig. 8).

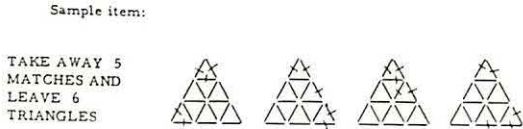


FIG. 8

The matches to be taken away are indicated by crossing them out with short heavy marks. In the sample item above the solution is reached by removing 5 matches in the four different patterns indicated.

Scoring: Credit is given for each correct, different pattern. The total score is the number of acceptable solutions.

Parts, 2; items per part, 5; working time, 20 minutes; maximum score, 38.

c. Symbolization

Test 18. Symbol Production—PEO6A. In this test *E* must produce symbols of his own invention to represent given activities and objects. In order to be acceptable the symbol produced must be a meaningful one for the action and the thing presented (see Fig. 9).

E is instructed not to draw stick figures or cartoon figures performing the

acts but to try to symbolize the actions themselves.

Scoring: Obviously a wide variety of symbols may be produced in this test. A response is judged as an acceptable symbol if some relation, however remote, can be seen between it and the object or action to be symbolized. Cartoon figures are not acceptable unless there is additional material that can stand alone as a proper symbol. The same symbol used for different items is scored only on its first occurrence. The total score is the number of acceptable symbols.

Parts, 2; items per part, Part 1-31, Part 2-30; working time, 10 minutes; maximum score, 61.

Test 19. Line Drawing—PEO7A. This is a test developed by J. P. and Ruth B. Guilford as a prognostic test of creativity in design (7). *E* is given a list of adjectives for each of which he is to draw a simple line that expresses the meaning of the adjective. Any kind of line may be drawn, such as straight, curved, in the form of a wave, or an angular pattern. The test is not a speeded one; everyone is allowed to finish.

Scoring: Each line is scored for *form*, *direction*, and *type*. Lines are described as to *form* in four ways: wave, curve, angle, and straight line. The *direction* is scored in three ways: horizontal, up (upward slant), and down (downward slant). The *type* may be heavy, medium, or light. This classification gives ten categories within which to classify each line in the test, i.e., four for *form*, three for *direction*, and three for *type*. Using the responses from half of the examinee group, the frequency with which each line fell into each of the 10 categories was tabulated. The frequencies were translated into sigma values above and below the mean and these values, adjusted by a shift of zero point to make them all positive, were used as the weights. The score

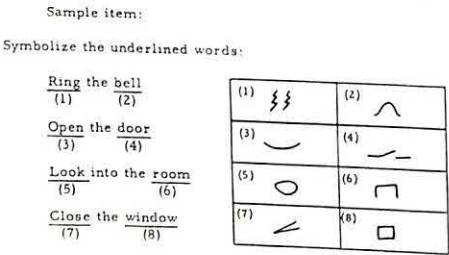


FIG. 9

for each line is the sum of the weights for form, direction, and type. The total test score is the sum of the line scores.

Parts, 1; items per part, 24; working time, 10 minutes; maximum score, 242.

4. Ordering

a. Temporal Ordering

Test 20. Temporal Ordering—*POO1A*.

In this test *E* is given a short description of a project and a list of some of the steps involved in the project. The steps are not in the best time order as listed. Following the list are questions that concern the proper order of the steps in different parts of the list. *E* indicates the answers by filling in the letters corresponding to the appropriate steps.

Sample item:

Fixing a Flat Tire:

While you are driving along a busy highway late one dark night, your right rear tire goes flat. The steps described below are some of those to be performed.

- Set out flares that had been stored in glove compartment.
- Tighten nuts on wheel.
- Block wheels so car won't roll.
- Raise car with bumper-jack.
- Replace flat with spare.
- Unlock the trunk to get wheel blocks.
- Take bumper-jack out of trunk.

In the blanks left in the following statements, enter the letters indicating the appropriate step or steps.

- The steps that should precede step *d* include:
a, c, f, g.
- The first two steps of this plan, in order, are:
a, f.
- The last step of this plan should be step *b.*

Scoring: The examinee had to decide, in general, which elements preceded and which followed the reference elements for a particular question. A weight of +1 is given each element properly placed. To correct for "padding," against which the examinees were forewarned, a weight of -1 is given each element incorrectly placed. Where the proper order was requested, bonus points are added

for responses in which the elements within a response are correctly ordered; points are deducted if the examinee's order is too far from the correct sequence. The score is the sum of the weights over all elements of the responses.

Parts, 4; items per part, 8; working time, 16 minutes; maximum score 75.

Test 21. Picture Arrangement. This is Dorothy C. Adkins' test based on sequences from the comic strip known as "Louie," which has been adapted with the permission of Press Features, Inc. Each item consists of a separate sequence in which the pictures are disarranged. *E* is required to number the pictures to produce what he considers the most sensible order (see Fig. 10).



FIG. 10

To show how the pictures should be arranged in order to make sense, a "1" has been written under the picture that should come first, a "2" under the picture that should come second, and so on.

Scoring: The total score is the number of correctly ordered sequences, i.e., items.

Parts, 2; items per part, I-8, II-12; working time, 7 minutes; maximum score, 20.

Test 22. Sentence Order—*RPO4B*. This was adapted from a test in Dorothy C. Adkins' North Carolina battery of reasoning tests. Each item consists of three sentences, which are to be arranged in a sensible order. *E* indicates the order by numbering the sentences.

Sample item:

- She bought some food at the market.
- She returned home and cooked some of the food she had bought.
- She went to the market.

Scoring: The total score is the number of items in which the sentences are correctly ordered.

Parts, 2; items per part, 10; working time, 6 minutes; maximum score, 20.

b. Hierarchical Ordering

Test 23. Outlining I—POO₂A. In this test a series of statements of differing degrees of generality are given and *E*'s task is to arrange the given statements so that the more particular (less general) statements are grouped under the more general statements with which they have most in common. One statement, the most general of all, is to be designated as the title.

Sample item:

- A. Cats are useful as mousers.
- B. Dogs are fine as guardians.
- C. Cows give milk.
- D. Animals are useful in a variety of ways.
- E. Some pets are useful as well as being enjoyable.
- F. Sheep furnish wool.
- G. Farm animals are essential to our way of life.

Instructions: Put the letter for each of the statements above in one of the spaces below so as to make the best outline. Use all statements; use each statement only once.

```

Title D
  I. G
    C
    F
  II. E
      A
      B
  
```

In the test proper the context of the hierarchical ordering task involves the characteristics of eggs as a food.

Scoring: Weights are assigned to each element on the basis of its being grouped with the other elements with which it properly belongs and on the basis of its being at the proper level of generality as intended by the constructors of the item. Score is the sum of weights over all elements of the response.

Parts, 1; elements per part, 16; working time, 5 minutes; maximum score, 21.

Test 24. Outlining II—POO₂A. This test was administered as Part II of the Outlining Test, for which the description appears above (test 23, Outlining I). The context of the hierarchical ordering task is the planning of a club meeting.

Scoring: See test 23.

Parts, 1; elements per part, 16; working time, 5 minutes; maximum score, 21.

Test 25. Word Matrices—CYO₅A. In this test *E* is given several word squares, each of which is composed of three horizontal spaces and three vertical spaces. Two or three of the spaces have already been filled in with words which set the pattern for filling in the rest of the spaces. *E*'s task is to arrange the words given below each square in their proper places so that the words of each row of the square will show the same relationships.

Sample item:

minnow	—	net
—	—	rod
—	—	—

lake, whale, ocean, bass, harpoon, pool.

The proper arrangement is:

minnow	pool	net
bass	lake	rod
whale	ocean	harpoon

The pattern in each of the horizontal lines is: type of fish, place it is usually found, and instrument with which it is caught. Each of the vertical columns is arranged so that all the items in each column are of a similar nature. For example, all the items in the first column are names of fish.

Scoring: The total score is the number of matrices in which all words are correctly arranged.

Parts, 1; items per part, 12; working time, 8 minutes; maximum score, 12.

5. Ingenuity

a. New Methods

Test 26. Unusual Methods—PIO₃A.

Each item involves a problem which is ordinarily handled in ways that are familiar to everyone; *E*'s task is to find two different and unusual ways of dealing with the problem. One restriction is that the ideas must be usable ones, although they need not be the best or as efficient as the methods already in use.

Sample item:

Suggest two unusual methods for accomplishing the following:

1. Relieve the boredom and fatigue of doing steady work in a business or industry. (The usual method is to have regular breaks during which employees relax or take refreshments).

Possible answers: *Have employees change their tasks every hour.*

Have four-hour shifts instead of eight-hour ones.

Scoring: The total score consists of the number of acceptable responses. In order for a method to be acceptable it must be reasonably practical and not now in use to any appreciable degree.

Parts, 2; items per part, 5; working time, 12 minutes; maximum score, 20.

Test 27. Verifications—PIO₂A. Certain phenomena of a physical or a biological nature are presented as facts to be verified. *E*'s task is to find ways of demonstrating these facts. The phenomena are stated in nontechnical terms and satisfactory answers may also be phrased in general terms. For each phenomenon given, *E* is required to write down two different procedures for demonstrating or investigating its validity.

Sample item:

Give two ways of demonstrating that the inside of a large gas flame is not as hot as the outside parts.

Possible answer: *Hold stick in flame—outside parts burn sooner.*

Put thermometer in center of flame and then in outside part—temperature rises faster in second part.

Scoring: The responses are weighted 1 if the general method suggested is an acceptable one. An additional weight of 1 or a total weight of 2 is given to responses that are complete and sound. The total score is the sum of the weights.

Parts, 2; items per part, 5; working time, 20 minutes; maximum score, 40.

b. New Applications

Test 28. Procedure Applications—PIO₁A. In each part *E* reads a description of a procedure as it is used in accomplishing some particular task. A generalized statement of the procedure then follows in order to make sure that all examinees know the procedural principle that is used. *E* must then find four other tasks or instances where the same procedure would apply.

Sample item:

In chemistry, one of the methods for getting a pure sample of some substance is by successive crystallizations. The substance is first put into solution, then crystallized out. This removes some impurities. The process is repeated several times; each crystallization removes additional impurities until only the pure substance is left. If we think of this method in the general sense of refining or purifying by repeating a process, in what other instances might this method be used?

Possible answers: *Athletic contests in which successive races or matches are used for elimination purposes.*

Spelling bee—people who misspell words drop out until only the best speller remains.

Rinsing clothes—each rinse takes out more soap until none remains.

Scoring: Each procedure has two major aspects. A weight of 2 is given for each response in which both aspects are applicable; a weight of 1 is given for each response in which one is applicable; and a weight of zero is given for each response in which neither aspect is applicable. The total score is the sum of the weights.

Parts, 5; items per part, 1; working time, 12.5 minutes; maximum score, 40.

6. Evaluation

a. Importance of Variables

Test 29. Essential Operations—RPO₃B.

E is provided with a problem and five items of information, one of which is irrelevant to the solution of the problem. *E* indicates which is the irrelevant statement, but does not work the problem. This is a revised form of the test used in previous reasoning studies (3, 6), incorporating certain revisions and additions in the items that were intended to reduce the general-reasoning variance and maximize the judgmental aspects of the test.

Sample item:

Question: How many miles apart will the two boats be when they run out of gas?

- A. Boats X and Y each have 5 gallons of fuel.
- B. They cruise together toward the same point.
- C. Boat X uses 1 gallon of gas going 3 miles.
- D. Boat Y uses 2 gallons of gas going 1 mile.
- E. Boat Y is twice as heavy as boat X.

The correct answer is *E*, since all the other statements, *A*, *B*, *C*, and *D*, provide necessary information for finding the answer to the question. The fact that boat Y is twice as heavy as boat X does not help to answer the question.

Scoring: $R - W/4$.

Parts, 2; items per part, 10; working time, 16 minutes; maximum score, 20.

Test 30. Ranking of Variables—PJO₂A.

This is an adaptation of the Awareness of Variables test described previously as test 5. Each item consists of a problem in which some decision has to be made and a number of facts that would be of aid in making the decision. *E*'s task in this test is to rank all the facts in order of their relative importance in reaching a decision.

Sample item: The sample item given for test 5 is also used here. This time the statements are ranked in order of importance with 1 indicating the most important and 5 marking the least im-

portant fact to consider. The ranks assigned to the statements in the sample item of test 5 are: A-4, B-3, C-2, D-1, E-5.

Scoring: The score for each item is determined from the sum of the differences obtained between the keyed ranks and the ranks assigned by the examinee. By this system, the greatest possible error for each item composed of seven elements is 24 points. (The maximum error would occur, for example, if the assigned rank order is the complete reverse of the keyed order). In order to make the higher scores correspond to the better responses, the sum of the obtained differences for each item is subtracted from 24.

Parts, 2; items per part, 2; working time, 10 minutes; maximum score, 96.

b. Seeing Deficiencies

Test 31. Seeing Deficiencies—PJO₁A.

In this test *E* is presented with short descriptions of plans or activities that for some reason will not lead to the desired result. *E*'s task is to point out in what way the plan is faulty.

Sample item:

A growing city discovers pressing needs to improve both its streets and its sewer system. After due consideration, the council decides to work on the street improvement program first. What is wrong with this plan?

Possible answer: *Improving streets before sewer system would require the street improvement job be done twice.*

Scoring: One written response is called for in each item and this must correspond in meaning to the keyed answer. No penalty is made for difference in expression of the answer. The total score is the number of items correctly answered.

Parts, 2; items per part, 10; working time, 20 minutes; maximum score, 20.

7. General Planning

Test 32. Planning Skills—PGO₁A. This

is a test by Irving Lorge used in studying decision-making. The test requires *E* to devise a complete plan for meeting a practical problem. The problem is similar to some of those encountered by the military in connection with human relations and morale. It involves the effect of the conditions at some hypothetical Air Force base in the Southwest upon the personnel assigned to the base. These conditions include large work load, lack of adequate working facilities, lack of recreational facilities, lack of facilities for dependents and relative isolation of the base. *E*'s task is to recommend a plan of action to solve the problems of low morale and low operating efficiency of the base.

Scoring: The procedure used by Lorge in scoring the test is employed here. Lorge had defined, by means of a content analysis, eight areas in which the responses could be grouped. Within each area there are general statements listed that serve to classify the responses as to quality and determine the weights that

should be attached to them. These weights were originally obtained by using a large sample of responses and demonstrating a high agreement among raters in setting the weights. The total score is the sum of the weights.

Parts, 1; items per part, 1; working time, 15 minutes; maximum score, U.

B. DESCRIPTION OF REFERENCE TESTS

1. Planning (AAF Analyses)

Test 33. Planning Air Maneuvers—CI408AX3. This test assumes that *E* is a sky-writing pilot who must plan how to write two adjacent letters by flying the shortest possible path. The starting and finishing positions of the plane are shown, and the sharpest turn that the plane can make is indicated. With this information, and the large letters to be written presented in the test, *E* must select the most direct path in terms of the instructions and indicate the direction he is traveling at each of the numbered points (see Fig. 11).

Sample item:

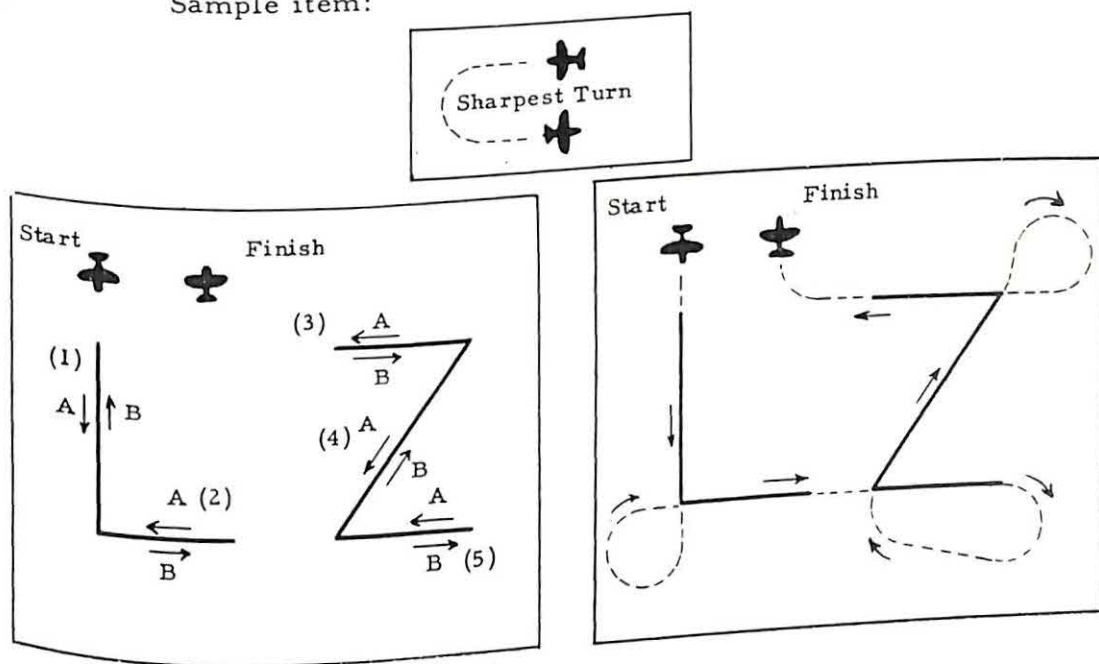


FIG. 11

By referring to the right-hand half of Fig. 11, it can be seen that the correct direction for 1 is A, for 2 is B, 3 is A, 4 is B, and 5 is B.

Scoring: Since the choice of a direction at one of the numbered points affects the direction choice at subsequent decision points in the same pattern, it was decided to score the test on the basis of total patterns. Within each pattern of letters, all choices have to be correct before credit is given. The score is the number of correct patterns.

Parts, 2; items per part, 10; working time, 16 minutes; maximum score, 20.

Test 34. Planning a Circuit—CI4OI. This is another Air Force test of planning. Each item consists of an electrical-circuit diagram with many intersecting wires and with several sets of terminals. *E*'s task is to trace the circuits and determine at which pair of terminals a battery should be placed in order to complete the circuit through a meter (see Fig. 12).

Sample item:

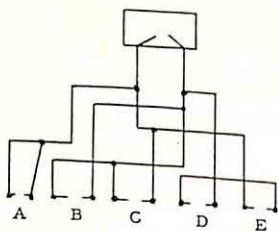


FIG. 12

Dots in the figure represent connections; that is, two wires are joined at that point. Where there is no such dot the insulated wires simply cross each other with no connection. Both poles of A are connected to the same point. This shorts out the battery, and will not make the meter work. B would also short out for the same reason. One pole of C is connected to one side of the meter and the other pole to the other side of the meter. C is therefore the correct answer.

Scoring: $R - W/4$.

Parts, 3; items per part, 12; working time, 12 minutes; maximum score, 36.

2. Integration III

Test 35. Code Analysis. Each item consists of a key row of five letters and numbers and below it five other similar rows which constitute the choices. *E*'s task is to select the choice row that contains a given number of symbols of the key row. The numbers 1 to 5 are interchangeable with the letters A to E, respectively. In other words, A is the same as 1, B is the same as 2, etc. The first number, five or less (or its equivalent letter), in the key row indicates the number of the symbols in the key row that the choice row should contain.

Sample items:

I. F E 5 3 A

A. H 3 3 3 A
B. 3 3 5 H 3
C. H C 1 E D
D. F 1 E E 3
E. 1 3 3 H D

II. J G 9 A 3

A. 9 7 9 1 3
B. J 6 1 C 7
C. G 7 B D 4
D. 1 5 D C F
E. 2 7 J E 1

Answers: I-D,

II-C.

Scoring: $R - W/4$.

Parts, 2; items per part, 20; working time, 22 minutes; maximum scores, 40.

3. Verbal Comprehension

Test 36. G-Z Verbal Comprehension. This test is Part I (Short Form) of the Guilford-Zimmerman Aptitude Survey. It is a standard vocabulary test, presenting in each item a given word and five alternative choices for the best synonym.

Scoring: $R - W/4$.

Parts, 1; items per part, 40; working time, 12 minutes; maximum score, 40.

Test 37. Vocabulary Test—CI6O4B. This is an Air Force test. It contains words of appropriate difficulty for men of approximately two years of college training.

Scoring: $R - W/4$, converted to standard value.

Parts, 1; items per part, 150; working time, 15 minutes.

4. General Reasoning

Test 38. Ship Destination—RGRO1B. Using a diagram and observing an increasing number of rules, *E* must determine which of three given ports is nearest to a ship in a given position. In addition to the rules of distance, the wind direction must be considered at the start of the test. Later the adequacy of the ports, the ocean current, and the initial heading of the ship must also be taken into account. All these conditions influence distance according to given rules.

Scoring: $R - W/2$.

Parts, 1; items per part, 15; working time, 15 minutes; maximum score, 15.

Test 39. Arithmetic Reasoning—GI2O6C. This is an Air Force test. The test consists of the setting up of an arithmetic solution to short problems. For example, if a plane uses a given amount of gasoline to fly a given distance, how much gasoline would be required to fly another given distance? In each item the correct response must be selected among five choices given.

Scoring: $R - W/4$, converted to stantine value.

Parts, 1; items per part, 30; working time, 35 minutes.

5. Logical Evaluation

Test 40. Logical Reasoning—RLRO5A. This is a syllogism-type test. The items include categorical, hypothetical, and mixed syllogisms, not all of which are strictly formal. Two, three, or four statements are given in an item and the examinee must decide whether the last one, or conclusion, follows from the preceding statements.

Sample item:

All bombers are heavier than air. All airplanes

are heavier than air. Therefore, all bombers are airplanes.

Answer: False.

Scoring: $R - W$.

Parts, 2; items per part, 15; working time, 10 minutes; maximum score, 30.

Test 41. Inference—RDO3B. Each item presents a statement of fact or opinion and five alternate conclusions. *E* indicates which of these can be drawn with most justification. The items are logically informal.

Sample item:

Most of the trees in the forest are green.

- A. There are no yellow trees in the forest.
- B. There are some yellow trees in the forest.
- C. Some of the trees in the forest are green.
- D. Green trees are the tallest in the forest.
- E. Pine trees are green.

Answer: C.

Scoring: $R - W/4$.

Parts, 2; items per part, 8; working time, 10 minutes; maximum score, 16.

6. Originality

Test 42. Plot Titles (Cleverness)—CFO2B. Four brief story plots are presented. *E* is instructed to write as many appropriate titles as he can for each plot.

Scoring: Two scores are derived from this test; a low-quality score and a cleverness score. The responses are rated clever or nonclever by comparing them with a comprehensive list prepared in connection with previous studies. The cleverness score is the total number of responses judged clever.

Parts, 4; items per part, 1; working time, 12 minutes; maximum score, U.

7. Ideational Fluency

Test 43. Plot Titles (Low Quality)—CFO2B. (See no. 42 above for description of test.)

Scoring: Total number of low-quality

Bureau Ednl. Resear.
DAVID H. ... COLLEGE

responses, i.e., responses not judged as clever.

Test 44. Consequences (Low Quality)—CFO6A. (See no. 8 for description of test.)

Scoring: Total number of low-quality responses.

8. Associational Fluency

Test 45. Controlled Associations II—CAFO2B. Each item consists of a common word for which *E* is to list three synonyms.

Sample item:

ODD: *unusual*
unique
queer

Scoring: Number of responses judged as sufficiently similar to the stimulus words.

Parts, 2; items per part, 10; working time, 6 minutes; maximum score, 60.

9. Adaptive Flexibility

Test 46. Match Problems—CXO3B. In this test groups of matches are laid out to form patterns of squares and triangles. *E*'s task is to take away a certain number of matches and have a certain number of squares or triangles remaining. They must come out even with no extra matches left over (see Fig. 13).

Sample item:

TAKE AWAY 2 MATCHES
AND LEAVE 2 SQUARES

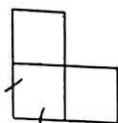


FIG. 13

The correct matches to be taken away are indicated by crossing them out with short heavy marks as has been done in the figure above. Notice that if any other pair of matches is taken away an extra match is left over in addition to the two squares.

Some of the test items are so constructed that they are insolvable if the subject imposes unnecessary restrictions upon himself, e.g., that the squares or triangles be the same size.

Scoring: Number of correct solutions.

Parts, 2; items per part, 9; working time, 12 minutes; maximum score, 18.

Test 47. Sign Changes—CXO4B. This test involves simple numerical operations in which *E* must substitute one arithmetic operation sign for another before performing the operation.

Sample item:

In the following problems, wherever you see:
— replace it by \times ; + replace it by $-$.

$$3 - 6 = 18$$

$$6 + 2 = -$$

$$4 - 3 = -$$

The answer for the second item should be 4 and for the third item 12.

Scoring: Number of correct responses.

Parts, 3; items per part, 8, 16, 16; working time, 2 minutes; maximum score, 32.

10. Education of Conceptual Relations

Test 48. Verbal Analogies I—RCRO1C. This is a standard verbal-analogies form of test. The only new idea is an attempt to make the relationship between the first pair of words relatively difficult to grasp and the analogy relatively easy to fulfill.

Sample item:

1. electricity : energy : : green : ?

A. color B. red C. grass D. trees E. foliage

2. athlete : scholar : : hands : ?

A. seclusion B. write C. study D. book E. mind

Answers: 1-A, 2-E.

Scoring: $R - W/4$.

Parts, 2; items per part, 15; working time, 12 minutes; maximum score, 30.

11. Judgment

Test 49. Practical Judgment—CI301C. This is an Air Force test. It requires the

selection of the best solution to a wide range of problem situations, mostly in a military setting; the items are of the type commonly called "work-planning" items. The answers are selected from groups of four or five alternatives.

Scoring: $2R - W/3$, converted to stanine value.

Parts, 1; items per part, 30; working time, 30 minutes.

12. Visualization

Test 50. Mechanical Principles—CI9O3B. This is an Air Force Classification test in which all items are presented pictorially. The task consists in the application of mechanical principles to the pictured problems. *E* has to select the correct response among three to five choices. For example, two ways of extracting a nail from a board with a claw hammer are pictured, and the problem is to determine which of the two ways, if either, requires more force.

Scoring: $R - W/2 + 20$, converted to stanine value.

Parts, 1; items per part, 40; working time, 20 minutes.

13. Symbol Manipulation

Test 51. Symbol Manipulation. The items consist of a main statement composed of letters and symbols followed by several conclusions also stated in symbols. On the basis of the symbols given, *E* determines whether each of the conclusions is true or false in relation to the main statement.

Sample item:

"L" means "larger than"

"S" means "smaller than"

"E" means "equal to"

"NL" means "not larger than"

"NS" means "not smaller than"

"NE" means "not equal to"

If $x S y$ (x is smaller than y), then:

1. $x E y$

2. $x NL y$

3. $x L y$

Answers: 1. False 2. True 3. False

Scoring: $R - W$.

Parts, 1; items per part, 25; working time, 7 minutes; maximum score, 25.

14. Numerical Facility

Test 52. Numerical Operations—CI7O2B. This is an Air Force Classification test consisting of two parts: Part I—deciding whether answers to simple addition and multiplication problems are right or wrong; Part II—finding the answer to simple subtraction and division problems. There are five choices.

Scoring: $R - 3W$, converted to stanine value.

Parts, 2; items per part, I-97, II-78; working time, I-5 minutes, II-5 minutes.

C. THE TEST ADMINISTRATION

The planning battery was administered to entering aircrew trainees at Lackland Air Force Base, Texas, in January 1954. The sample number for which scores were obtained on all the tests was 364. The test administration was reported to have run smoothly and the examinees were assumed to have worked at the tests with fairly high motivation. The Aircrew Classification Battery had been administered just prior to the experimental battery.

IV. TREATMENT OF THE DATA

The classification tests were scored by the Air Force and scores converted to stanine values; the remaining tests were scored at this laboratory.

The test reliabilities were computed on a subsample of 200 examinees. The values and types of the reliability estimates are given in Table 1 along with

TABLE 1
MEANS, STANDARD DEVIATIONS, AND
RELIABILITIES OF SCORES

Test	M	σ	r_{11}	Est. ^a
1. Matrix Order	14.6	4.2	.70	AF
2. Seeing Trends	9.7	4.5	.81	AF
3. Sensitivity to Order	13.0	2.1	.36	AF
4. Pertinent Questions	23.5	4.4	.80	AF
5. Awareness of Variables	9.9	2.1	.44	AF
6. Series	9.5	2.0	.44	KR
7. Effects	32.0	10.0	.84	AF
8. Consequences (remoteness)	9.7	4.4	.33	AF
9. Competitive Planning	19.1	6.8	.58	AF
10. Symbol Grouping	17.3	4.0	.65	AF
11. Contingencies	15.6	4.3	.60	AF
12. Route Planning	31.3	6.4	.77	AF
13. Planning Skills II	11.3	4.0	.73	AF
14. Planning Elaboration	15.5	4.1	.57	AF
15. Figure Production	23.9	7.7	.77	AF
16. Alternate Methods	11.4	2.5	.50	AF
17. Match Problems II	17.1	4.8	.65	AF
18. Symbol Production	26.2	8.6	.86	AF
19. Line Drawing	188.3	13.4	.63	AF
20. Temporal Ordering	37.5	11.8	.67	AF
21. Picture Arrangement	15.9	2.5	.60	AF
22. Sentence Order	11.4	2.6	.56	AF
23. Outlining (Part I)	12.6	3.5	.57	KR
24. Outlining (Part II)	10.6	3.8	.50	KR
25. Word Matrices	6.8	2.5	.79	OE
26. Unusual Methods	10.1	3.5	.74	AF
27. Verifications	18.3	6.2	.72	AF
28. Procedure Applications	10.5	4.8	.63	AF
29. Essential Operations	11.6	4.0	.62	AF
30. Ranking of Variables	59.8	8.8	.45	AF
31. Seeing Deficiencies	9.9	2.8	.58	AF
32. Planning Skills	13.0	4.0	.38	1 ²
33. Planning Air Maneuvers	5.2	2.4	.48	AF
34. Planning a Circuit	20.0	5.6	.73	AF
35. Code Analysis	15.0	7.5	.71	AF
36. G-Z Verbal Comprehension	20.6	6.2	.81	OE
37. Vocabulary	5.0	1.8	.83	TR-A
38. Ship Destination	6.5	3.3	.60	KR
39. Arithmetic Reasoning	5.6	1.7	.70	OE-A
40. Logical Reasoning	10.0	6.2	.58	AF
41. Inference	6.4	3.6	.67	AF
42. Plot Titles (cleverness)	6.4	3.7	.43	AF
43. Plot Titles (low quality)	12.2	6.9	.70	AF
44. Consequences (low quality)	13.1	4.8	.40	AF
45. Controlled Associations II	21.6	7.2	.67	AF
46. Match Problems	7.9	3.0	.56	AF
47. Sign Changes	21.6	4.5	.69	AF
48. Verbal Analogies I	12.3	4.4	.54	AF
49. Practical Judgment	5.4	1.9	.43	OE-A
50. Mechanical Principles	5.4	1.9	.82	OE-A
51. Symbol Manipulation	15.1	4.1	.48	KR
52. Numerical Operations	5.3	1.7	.81	AF-A

^a AF: alternate-form estimate; OE: odd-even estimate; Spearman-Brown formula or Horst formula (where parts contained unequal numbers of items) was applied in AF and OE cases, except for tests 8, 42, 43, and 44; KR: Kuder-Richardson estimates; TR: Test-retest, A as a third letter refers to Air Force Classification tests were computed from stanine scores. The communality value is used as the lower-bound estimate of the reliability of test 32 because of the impossibility of making a valid estimate from what is essentially a single-item test.

the means and standard deviations of the tests.

Score distributions for most of the variables did not appear to be signifi-

cantly deviant from normal and the raw scores were used in computing Pearson r s. The distributions of four of the variables (tests 12, 21, 42, 43) were markedly skewed. For these variables normalized scores were used in computing Pearson r s.

Two scores each were derived from the Plot Titles and Consequences tests, a low-quality and a cleverness score for the former, and a low-quality and a remoteness score for the latter. In the correlation of one score of a test with the other score of the same test, the two cross correlations of separately timed part scores were computed and averaged in order to arrive at a correlation value relatively uncontaminated by any experimental dependence of the two scores. For all other correlations in which these tests (Plot Titles and Consequences) were involved, the correlation coefficient used in the analysis was an average of those from two part scores. The correlation matrix is given in Table 2.

Seventeen centroid factors were extracted.³ The highest coefficient in each column of the correlation matrix was used as the estimate of the communality of the corresponding variable. After each extraction the highest residual value in each column was used as an estimate of the communality remaining. The loadings of the last three centroid factors were all smaller than .20 but these factors were used in the later rotations in order to help clarify the other factors.

The Zimmerman graphic, orthogonal

³ A copy of the centroid factor matrix has been deposited with the American Documentation Institute, Order Document No. 5269 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D.C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 × 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

method was used in rotating the axes (14). The criteria for rotations in the early stages were those of simple structure and positive manifold. Where these criteria were in conflict, the rotation was guided by the criterion of simple structure. After each axis had been rotated at least once, psychological meaningfulness and agreement with previous well-known factor-analytic results became important criteria. The final rotations were adjustive in nature in order to improve simple structure and positive manifold. The final rotated factor loadings are presented in Table 3.

V. INTERPRETATION OF THE FACTORS

The factors are presented in the approximate order of their familiarity and definiteness. The interpretations rest principally upon those tests with loadings of .30 or greater. The test numbers, names, and factor loadings are listed preceding the discussion of each factor. In parentheses are loadings .30 or greater on other factors. Discussion of the reference factors is omitted except where the factor is relatively less well-known.

Factor A. Verbal Comprehension (V)

37 Vocabulary	.73
36 G-Z Verbal Comprehension	.71
41 Inference Test	.32 (.30 LE, .30 Or)
45 Controlled Associations II	.30

Factor B. Numerical Facility (N)

52 Numerical Operations	.66
47 Sign Changes	.51
39 Arithmetic Reasoning	.43 (.31 GR)

Factor C. Visualization (Vz)

50 Mechanical Principles	.44 (.32 AX)
12 Route Planning	.32 (.38 PF, .34 AX)

Factor D. General Reasoning (GR)

38 Ship Destination	.44
35 Code Analysis	.35 (.33 PF)
39 Arithmetic Reasoning	.31 (.43 N)
29 Essential Operations	.31 (.31 LE, .31 J)
23 Outlining (Part I)	.31

Factor E. Logical Evaluation (LE)

40 Logical Reasoning	.46
29 Essential Operations	.31 (.31 GR, .31 J)
41 Inference Test	.30 (.32 V, .30 Or)

Factor F. Ideational Fluency (IF)

44 Consequences (Low Quality)	.49
43 Plot Titles (Low Quality)	.37

The two tests listed for factor F were included in the battery primarily as reference tests for *ideational fluency*. The factor was previously defined (13) as the speed of calling up ideas in a situation in which there is relatively little restriction, and quality does not matter.

Factor G. Judgment (J)

31 Seeing Deficiencies	.46 (.36 Or)
11 Contingencies	.43 (.39 CF)
27 Verifications	.38
29 Essential Operations	.31 (.31 GR, .31 LE)
49 Practical Judgment	.30
14 Planning Elaboration	.30 (.44 E)

The *judgment* factor usually appears whenever "judgment" tests are included that are composed of items presenting a predicament and requiring the selection of the best solution. In the AAF analyses (9) the hypothesis offered for this factor was "the ability to weigh solutions and select the wisest and best one."

These results indicate that *judgment* is not bound to the multiple-choice test format. The individual himself can produce solutions prior to weighing the merits of solutions. A fluent recall of experiences that are relevant to the problem situation would provide the individual with a wide range of possible solutions from which to choose. In examining a number of solutions to a practical problem it is not unlikely that the best choice can be made when the examinee considers the given solutions against the background of his experience, placing them in a hierarchy of solutions that he produces.

TABLE 2
THE CORRELATION MATRIX*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
253	253	180	121	166	191	117	161	175	079	214	199	050	123	152	189	193	271	055	296	262	271	218	125	329
180	137	137	074	072	100	068	068	088	054	152	267	041	138	090	107	187	299	168	251	157	251	216	216	243
121	232	074	170	170	132	626	340	062	036	482	163	467	465	257	516	136	316	023	214	160	219	065	223	141
166	138	072	170	105	123	108	043	018	132	048	136	104	064	085	055	113	113	144	181	102	117	085	130	130
191	264	100	132	105	162	167	167	128	105	146	049	130	166	220	195	263	128	256	206	201	151	131	191	283
117	273	068	626	123	162	354	093	031	422	071	488	464	371	542	089	340	001	199	133	202	094	281	194	194
161	219	068	340	108	167	354	121	050	333	066	292	345	214	342	038	248	082	164	122	188	128	211	109	109
175	171	088	062	043	167	093	121	153	092	232	039	101	052	136	303	116	062	156	152	093	113	068	108	078
079	065	054	036	018	128	050	153	088	160	062	037	024	107	170	024	-004	018	096	087	059	055	055	055	055
214	266	152	482	132	105	422	333	092	088	055	354	372	230	398	138	279	068	245	092	292	140	285	252	252
199	132	267	163	048	146	071	066	232	160	055	-065	039	016	121	372	134	128	248	266	110	094	010	202	202
050	121	041	467	136	049	488	292	039	062	354	-065	479	287	336	-032	117	-014	093	023	168	095	200	174	174
123	226	138	465	104	130	464	345	101	037	372	039	479	296	463	086	231	018	177	130	171	104	274	106	106
152	144	090	257	064	166	371	214	052	024	230	016	287	296	290	056	280	024	156	204	163	056	172	106	106
189	223	107	516	085	220	542	342	136	107	398	121	336	463	290	204	314	035	175	137	169	114	189	189	189
193	173	187	136	055	195	089	038	303	170	138	372	-032	086	056	204	100	-012	182	231	119	092	022	181	181
271	299	119	316	113	263	340	248	116	024	279	134	117	231	280	314	100	132	310	151	201	166	177	191	191
055	168	047	023	113	128	001	082	062	-004	068	128	-014	048	024	035	-012	132	148	084	177	-010	061	303	303
296	251	298	214	144	256	199	164	156	048	245	248	093	177	156	175	182	310	148	261	323	154	186	303	303
262	157	156	160	181	206	133	122	152	096	092	266	023	130	204	137	231	151	084	261	259	149	148	303	303
271	251	233	219	102	201	202	188	093	087	292	110	168	171	163	169	119	201	177	323	259	166	187	343	343
218	169	164	065	117	151	094	128	113	059	140	094	095	104	056	114	092	166	-010	154	149	166	200	200	200
125	216	085	223	085	131	281	211	068	055	285	040	200	274	172	189	022	177	061	186	148	187	200	200	200
329	248	336	141	130	285	194	190	106	078	252	202	164	174	106	187	185	196	151	305	304	342	146	204	204
134	223	053	409	014	086	459	300	061	134	303	071	344	331	304	356	-001	292	014	120	099	086	147	144	047
242	285	169	376	186	202	421	292	060	-010	397	145	290	323	220	339	104	206	062	242	082	172	144	196	256
234	266	171	338	194	198	437	282	034	-067	325	050	266	363	238	346	067	248	115	228	208	334	143	196	343
264	254	227	158	269	204	213	222	198	067	271	187	120	191	029	213	215	105	084	268	136	282	163	250	343
088	117	134	051	145	102	-006	063	035	-005	074	049	030	029	-079	045	088	050	089	093	123	058	087	088	194
358	287	178	375	230	185	296	240	103	076	350	168	202	282	181	333	170	212	032	281	255	364	226	266	303
105	191	149	290	053	099	340	218	-034	044	306	-008	302	347	267	262	-032	266	023	149	082	257	170	133	161
148	110	158	018	151	144	032	-016	189	057	016	240	-034	016	016	098	284	114	-005	226	080	155	090	090	153
120	120	152	153	037	206	151	104	194	118	094	317	037	102	084	142	256	096	002	086	224	109	020	-009	261
230	112	146	018	110	192	058	110	294	178	044	215	035	108	-021	152	303	080	057	221	147	171	248	132	296
276	294	225	188	276	186	248	200	035	-045	301	096	145	200	081	163	052	172	189	328	124	389	155	229	296
286	336	218	209	219	137	243	216	049	-050	404	056	170	198	151	142	010	235	211	369	126	416	135	220	345
203	248	230	104	148	319	119	166	258	072	151	252	105	123	045	046	181	084	174	191	264	265	192	123	296
235	278	273	076	200	128	143	154	171	002	170	260	060	107	038	175	251	124	020	255	129	201	266	200	296
210	283	230	162	238	203	177	160	120	019	122	245	044	178	051	208	188	066	065	238	146	262	103	147	296
316	314	195	253	226	160	288	230	116	026	332	191	158	195	190	254	181	217	127	324	255	342	148	180	354
100	232	115	253	134	127	288	291	080	-084	244	034	174	245	241	274	019	276	088	184	092	160	126	100	161
-012	022	-103	286	-076	016	294	140	081	114	158	-026	270	284	178	204	015	098	016	-014	013	-034	034	010	203
009	062	022	252	034	029	304	240	-034	050	181	-039	243	248	205	244	082	140	032	036	013	020	020	186	219
196	261	218	308	075	172	424	232	088	049	321	060	234	292	198	319	114	208	106	218	-004	234	113	186	219
162	165	183	098	089	254	087	078	277	115	047	395	-054	050	161	130	455	111	-010	228	263	108	133	082	219
115	213	075	085	-048	148	106	133	123	130	180	148	012	073	047	120	068	192	101	160	071	130	144	202	181
301	196	368	169	153	224	183	168	122	064	205	170	179	186	118	213	201	116	044	264	198	286	188	085	383
195	105	196	129	112	023	087	048	-027	-017	106	164	043	078	070	045	154	110	093	175	129	216	-006	056	163
127	127	249	070	132	136	028	-028	047	-016	008	249	051	-029	-047	040	152	100	-026	090	-089	039	028	011	124
160	047	139	040	048	115	072	058	114	003	-035	105	-015	059	-004	086	115	035	-021	177	099	168	128	008	142
084	142	071	102	007	075	135	196	188	123	152	150	094	138	017	132	184	068	042	152	031	118	081	234	142

Decimal points have been omitted.

Factor H. Eduction of Conceptual Relations (CR)

48 Verbal Analogies I	.47
3 Sensitivity to Order	.46

This factor emerged in the two reasoning studies (3, 6). In both of those analyses, Verbal Analogies I was the leading test. The test was specifically designed to emphasize the discovery of

meaningful relationships, and the interpretation of the factor was based upon this intent.

The discovery feature is certainly important in Sensitivity to Order also. In this test adjustments in the given orders of words are to be made when one of the words is seen to be out of order. Although the emergence of factor H in

TABLE 2 (Continued)

27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
242	234	264	088	358	105	148	120	230	276	286	203	235	210	316	100	-012	009	196	162	115	301	195	127	160	084
235	266	254	117	287	191	110	120	112	294	336	248	278	283	314	232	022	062	261	165	213	196	105	127	047	142
169	171	227	134	178	149	158	152	116	225	218	230	273	230	195	115	-103	022	218	183	075	368	196	249	139	071
376	338	158	051	375	290	018	153	018	188	209	104	076	162	253	253	286	252	308	098	085	169	129	070	040	102
186	194	269	145	230	053	151	037	110	276	219	148	200	238	226	134	-076	034	075	089	-048	153	112	132	048	007
202	198	204	102	185	099	144	206	192	186	137	319	128	203	160	127	016	029	172	254	148	224	023	136	115	075
421	437	213	-006	296	340	032	151	058	248	243	119	143	177	288	288	294	304	424	087	106	183	087	028	072	135
292	282	222	063	240	218	-016	104	110	200	216	166	154	160	230	291	140	240	232	078	133	168	048	-028	058	196
090	034	198	035	103	-034	189	194	294	035	049	258	171	120	116	080	081	-034	088	277	123	122	-027	047	114	188
-010	-067	067	-005	076	044	057	118	178	-045	-050	072	002	019	026	-084	114	050	049	115	130	064	-017	-016	003	123
397	325	271	074	350	306	016	094	044	301	404	151	170	122	332	244	158	184	321	047	180	205	106	008	-035	152
145	050	187	049	168	-008	240	317	215	096	056	252	260	245	191	034	-026	-039	060	395	148	170	164	249	105	150
290	266	120	030	202	302	-034	037	035	145	170	105	060	044	158	174	270	243	234	-054	012	179	043	051	-015	094
323	363	191	029	282	347	016	102	108	200	198	123	107	178	155	245	284	248	292	050	073	186	078	-029	059	138
220	238	029	-079	181	267	016	081	-021	081	151	045	038	051	190	241	178	205	198	161	047	118	070	-047	-004	017
339	346	213	045	333	262	098	142	152	163	142	046	175	208	254	274	204	244	319	130	120	213	045	040	086	132
104	067	215	088	170	-032	284	256	303	052	010	181	251	188	181	019	015	082	114	455	068	201	154	152	115	184
206	248	105	050	212	266	114	096	080	172	235	084	124	066	217	276	098	140	208	111	192	116	110	100	035	068
062	115	084	089	032	023	-005	002	057	189	211	174	020	065	127	088	016	032	106	-010	101	044	093	-026	-021	042
242	228	268	093	281	149	226	086	221	328	369	191	255	238	324	184	-014	036	218	238	160	264	179	090	177	152
082	208	136	123	255	082	080	224	147	124	126	264	129	146	255	092	013	013	-004	263	071	198	129	-089	099	031
172	334	282	058	364	257	155	109	171	389	416	265	201	262	342	160	-034	020	234	108	130	286	216	039	168	118
144	143	163	087	226	170	090	020	248	155	135	192	266	103	148	126	034	020	113	133	144	188	-006	028	128	081
196	196	250	088	266	133	090	-009	132	229	220	123	200	147	209	180	100	040	186	082	122	208	056	011	008	234
259	286	347	197	302	117	166	152	261	290	349	292	290	298	357	160	-087	014	206	219	184	382	183	068	124	142
280	232	063	-099	164	344	-040	149	041	028	055	113	-039	-002	176	186	252	216	188	077	069	-005	008	-027	-055	070
365	365	286	140	302	223	096	148	065	330	310	156	237	289	284	204	138	086	318	118	064	260	147	305	031	-016
286	229	098	088	380	282	065	122	041	324	356	167	234	280	309	250	169	150	378	086	096	186	225	032	132	122
140	098	088	088	376	119	179	097	286	358	343	325	371	402	288	102	-009	-032	180	154	169	356	245	188	216	226
302	380	376	089	089	039	059	008	142	178	130	193	144	169	126	068	-134	102	096	034	063	121	154	117	084	049
223	282	119	039	257	188	232	155	399	381	285	281	284	284	374	132	004	048	304	156	063	334	299	106	122	086
066	065	179	059	188	-016	062	-001	217	249	133	047	059	171	179	148	223	259	007	093	128	105	-103	073	021	32
148	122	097	008	232	062	215	240	097	074	172	060	-005	130	156	054	-088	-068	060	378	122	181	093	240	109	060
065	041	286	142	155	-001	240	110	090	050	302	325	238	152	070	038	-048	052	233	220	282	-013	008	122	310	35
330	324	358	178	399	217	097	078	090	782	782	243	327	325	447	144	-109	-042	398	119	089	473	300	166	297	088
310	356	343	130	381	249	109	074	050	782	240	295	245	456	162	-074	-023	414	102	130	357	273	069	212	162	37
156	167	325	193	285	133	145	172	302	243	240	246	323	205	124	-074	-087	174	200	142	299	124	163	170	169	38
237	234	371	144	281	047	257	060	325	327	295	246	360	302	115	-048	-003	263	262	259	372	135	122	209	399	39
280	280	402	169	284	059	250	-005	238	325	245	323	360	370	114	-076	-034	147	240	106	379	178	238	192	129	40
284	309	288	126	374	171	156	130	152	447	456	205	302	370	187	031	032	244	180	155	326	194	055	147	164	41
204	250	102	068	132	179	054	026	070	144	162	124	115	114	187	058	088	187	109	099	129	014	-012	036	099	42
138	169	-009	-134	004	148	-088	-008	038	-109	-074	-074	-048	-076	031	088	292	195	-052	052	-117	-060	-153	-079	113	43
086	150	-032	102	048	223	-068	-002	-048	-042	-023	-087	-003	-034	032	088	187	188	021	090	052	-013	-048	-023	017	44
318	378	180	096	304	259	060	054	052	398	414	174	263	147	244	187	095	188	100	057	306	202	079	154	139	45
118	086	154	034	156	007	378	258	233	119	102	200	262	240	180	109	-052	021	100	048	247	128	176	114	120	46
064	096	169	063	063	093	122	078	220	089	130	142	259	106	155	099	052	090	057	048	093	000	-101	140	425	47
260	186	356	121	334	128	181	084	282	473	357	299	372	379	326	129	-117	052	306	247	128	000	185	186	007	48
147	225	245	154	299	105	093	117	-013	300	273	124	135	178	194	014	-060	-013	079	176	-101	260	186	018	-002	49
305	032	188	117	106	-103	240	182	008	166	069	163	122	238	055	-012	-153	-048	079	176	-101	260	186	018	-002	50
021	132	216	084	122	073	109	093	122	297	212	170	209	192	147	036	-079	-023	154	114	140	228	007	018	118	51
-016	122	226	049	086	021	060	-007	310	088	162	169	399	129	164	099	113	017	139	120	425	089	-002	-220	118	52

this analysis lends confirmation to the existence of this ability, the definition has not been clarified, due to the paucity of tests with loading on the factor.

Factor I. Originality (O)

18 Symbol Production	.42
42 Plot Titles (Cleverness)	.36
2 Seeing Trends	.35
8 Consequences (Remoteness)	.34 (.30 CF)

26 Unusual Methods
15 Figure Production

.32 (.34 E)
.30 (.37 E)

This seems to be the *originality* factor that emerged in the analysis of creative-thinking tests. In that study, Plot Titles was the leading test on *originality*, and Consequences also had a significant loading.

Symbol Production was designed as a

TABLE 3
ROTATED FACTOR LOADINGS

	A V	B N	C Vz	D GR	E LE	F IF	G J	H CR	I O	J PF	K AX	L Or	M E	N CF	O (r)	P (r)	Q (r)	h ²
1.	.03	-.02	-.09	.11	.17	.00	.20	.22	.17	.12	.12	.33	-.12	.01	-.12	.06	.04	.39
2.	.15	.17	.14	.19	.11	.04	.23	.00	.35	.06	.11	.14	.02	.04	-.08	.08	-.07	.35
3.	.03	.11	.24	.03	.05	-.06	.10	.46	.02	.01	.09	.24	.12	-.05	.01	-.10	.07	.40
4.	.06	.03	.08	.02	.03	.24	.27	-.10	.11	.07	.09	.14	.28	.58	.00	-.02	-.17	.65
5.	.18	-.08	.02	.26	.24	.06	.09	-.06	.08	-.08	.13	.15	-.12	.10	.10	-.08	-.07	.28
6.	-.03	-.02	.07	.25	-.03	.00	.00	.12	.28	.18	.15	.24	.05	.02	-.02	.16	-.15	.33
7.	.04	.08	-.03	.09	.06	.26	.28	-.03	.17	.01	.07	.09	.47	.46	-.04	.13	-.13	.68
8.	.06	.11	-.01	.10	.14	.09	.14	.02	.34	.14	-.05	.06	.10	.30	-.01	-.02	.05	.34
9.	-.01	.16	-.01	.24	.01	-.01	-.03	.00	.06	.40	.18	.07	.02	.00	.03	.10	.03	.30
10.	-.08	.06	-.06	-.02	-.06	.08	.09	.08	.03	.36	.08	-.02	-.08	.00	.08	-.14	-.09	.22
11.	.19	.18	.03	.00	-.08	.03	.43	.02	.23	.08	-.06	.14	.13	.39	.00	-.06	-.01	.53
12.	-.02	.16	.34	-.06	.17	.00	.03	.13	-.05	.38	.34	.17	-.02	.00	-.15	-.02	-.02	.51
13.	.06	-.02	.07	.15	.02	.22	.23	-.03	.07	.05	-.14	.03	.40	.31	.10	-.16	.06	.49
14.	.00	.05	.05	.11	.12	.26	.30	-.04	.19	.12	-.09	.08	.44	.19	.21	-.03	-.03	.53
15.	-.02	-.06	-.09	-.09	.07	.14	.15	.02	.30	.00	.12	.17	.37	.05	.04	.02	.02	.34
16.	-.06	.08	-.07	.01	.13	.18	.24	.11	.18	.13	.10	.11	.26	.44	.00	.16	-.05	.51
17.	-.01	.18	.11	.07	.00	.14	.03	.09	-.10	.32	.43	.21	-.09	.05	.11	.10	.16	.47
18.	-.06	.09	.00	-.04	.00	.06	.10	.06	.42	-.01	.15	.24	.09	.15	.16	.10	-.10	.41
19.	.15	.04	.10	.02	.03	.09	-.06	-.01	.26	.07	-.14	.12	-.08	-.08	-.01	.05	-.13	.21
20.	.15	.19	.05	.07	.08	.02	.11	.18	.20	.06	.14	.36	.04	.09	-.02	-.01	-.17	.34
21.	.02	-.12	.03	.04	.19	.07	-.04	-.04	.09	.23	.12	.53	.03	.02	-.12	-.06	.10	.44
22.	.27	.10	.03	.06	.03	.00	.15	.17	.11	.07	.04	.45	.10	.03	-.01	.13	-.18	.42
23.	-.04	.12	-.15	.31	-.01	.00	.16	.13	.11	.02	.08	.14	.02	-.02	-.10	-.14	.06	.25
24.	.03	.21	-.04	.19	.10	.02	.26	-.05	.14	.01	-.09	.20	.12	.04	.12	.01	-.03	.26
25.	.11	.11	.16	.13	.12	-.07	.10	.26	.14	.14	-.04	.44	.02	.07	.07	-.01	-.02	.41
26.	-.08	-.02	-.06	-.04	-.01	.10	.28	-.06	.32	.16	.11	-.06	.34	.23	-.16	-.20	-.05	.49
27.	.08	-.04	.20	.23	.12	.00	.38	.13	.15	-.03	.10	.04	.23	.26	-.04	.12	-.06	.44
28.	.18	.03	.05	.12	.16	.20	.23	.04	.13	-.10	-.02	.32	.20	.17	-.15	.07	.01	.44
29.	.16	.16	.08	.31	.31	-.07	.31	.16	-.04	.18	.02	.14	.03	.06	.00	-.03	-.12	.45
30.	.08	.05	.23	.25	-.02	.10	-.04	.08	.09	-.09	.00	.13	-.24	.13	.06	-.08	.12	.28
31.	.21	-.02	-.05	.15	.15	.04	.46	.08	-.01	.09	.09	.36	.02	.17	-.06	-.05	.02	.50
32.	.11	.00	.07	.03	-.07	.22	.23	.13	.24	-.06	-.02	.07	.20	.00	-.12	-.26	-.03	.38
33.	.04	.18	.06	.14	.03	-.12	.05	.05	-.07	.02	.47	.19	-.02	.08	.14	-.02	-.13	.38
34.	.08	-.08	.22	.02	-.08	-.09	.07	.04	-.10	.39	.35	.21	.03	-.02	-.15	.03	.02	.44
35.	-.12	.25	-.13	.35	.14	-.05	-.01	.18	.00	.33	.10	.18	-.07	-.03	.12	-.07	-.02	.44
36.	.71	.04	.02	.20	.17	.00	.24	.28	.12	-.02	-.03	.14	.03	.02	-.03	.12	-.14	.79
37.	.73	.16	.00	.09	.04	-.03	.29	.17	.22	-.03	-.08	.21	.05	-.02	-.05	.08	-.09	.80
38.	.13	.01	.24	.44	.08	-.08	.00	.13	.11	.20	.00	.20	.06	-.01	-.14	-.17	.02	.49
39.	.11	.43	.00	.31	.25	.00	.16	.21	-.06	-.02	.15	.16	.00	-.01	-.04	.07	.14	.51
40.	.14	.14	.21	.28	.46	-.05	.04	.21	-.02	-.07	.14	.15	.03	.12	.04	-.01	-.10	.50
41.	.32	.12	.01	.03	.30	-.01	.18	.11	.17	.06	.11	.30	.04	.16	-.02	-.01	-.03	.42
42.	.01	.12	.01	.09	.05	-.07	.02	-.02	.36	-.10	.01	.16	.23	.20	.03	.03	.00	.35
43.	-.16	.08	-.12	-.03	.00	.37	.09	-.18	.04	.13	-.07	-.10	.27	.09	.01	.02	-.11	.37
44.	-.09	.02	-.02	-.08	-.04	.49	.01	.09	.16	-.06	.01	-.06	.08	.23	.09	-.06	.02	.45
45.	.30	.10	.02	.10	-.08	.23	.27	.28	.10	-.02	-.01	.07	.21	.16	-.01	.18	.07	.45
46.	.04	.12	.09	.07	.08	-.01	-.04	.09	.04	.23	.57	.10	.10	-.08	.10	.06	.15	.51
47.	-.04	.51	-.04	.03	.04	-.05	-.01	.09	.17	.12	-.04	.06	-.06	.02	-.20	-.03	-.12	.39
48.	.18	.05	.07	.23	.19	.04	.16	.47	-.01	.08	.04	.23	.05	.06	.18	.04	-.03	.48
49.	.22	-.03	.25	-.07	.12	.14	.30	.09	-.07	-.06	.06	.21	-.08	-.09	-.03	-.02	.06	.33
50.	-.04	-.12	.44	.16	.00	-.18	.23	.20	-.10	-.09	.32	-.10	-.08	.02	.03	.18	-.18	.56
51.	.22	.08	-.15	.18	.10	-.02	-.15	.28	-.10	.05	.09	.08	.03	.05	-.22	.03	-.04	.30
52.	.08	.66	-.10	.09	.17	.06	-.01	-.03	.01	.24	-.10	.01	.04	.00	-.01	-.04	.08	.57

test of symbolic thinking, along with the Line Drawing test (8). Both tests had their highest loading on *originality*, although in the case of the Line Drawing test, which had a very low communality, the loading is only near-significant (.26). Apparently the invention of symbols involves originality as much as those tests that call for clever, remote, and uncommon responses.

An interesting paradox may be noted in that the Line Drawing and the Symbol Production tests were scored on the basis of *commonness* (in the former test) and *intelligibility* (in the latter) of the symbols. Seemingly, a similar psychological process is demonstrated in one type of test resulting in uncommon responses, and in other tests resulting in responses that are common and readily understood.

Perhaps the conception of the factor of *originality* needs to be broadened to include symbolizing activities. The production of symbols of the kinds involved in tests 18 and 19 may be a sufficiently uncommon act in itself to require originality, so that the scoring can stress common responses and yet measure *originality*. In any event, Symbol Production emerges in this study as a very promising test of *originality*.

Factor J. Perceptual Foresight (PF)

9 Competitive Planning	.40
34 Planning a Circuit	.39 (.35 AX)
12 Route Planning	.38 (.34 Vz, .34 AX)
10 Symbol Grouping	.36
35 Code Analysis	.33 (.35 GR)
17 Match Problems II	.32 (.43 AX)

This factor seems to involve the ability to explore visually possible courses of action in order to select the most effective ones for getting solutions to detailed visual problems. Competitive Planning, Route Planning, and Symbol Grouping were selected or designed as measures of foresight. Each of these tests presents a task in which, ostensibly, several alternative courses may be pursued to the goal. Only one of the courses, however, is considered to fulfill the criterion given. In each case the examinee must look well beyond the step at which he is working in order to select the proper course of action. The steps or moves that are most obvious or seem to provide immediate gains are not always the best in terms of future gains or in leading to a correct solution.

Factor K. Adaptive Flexibility (AX)

46 Match Problems	.57
33 Planning Air Maneuvers	.47
17 Match Problems II	.43 (.32 PF)
34 Planning a Circuit	.35 (.39 PF)
12 Route Planning	.34 (.34 Vz, .38 PF)
50 Mechanical Principles	.32 (.44 Vz)

This has been identified as *adaptive*

flexibility, principally on the basis of Match Problems, which served as a reference test for this factor. The presence of three Air Force planning tests on this factor might suggest that this is the Air Force planning factor. It is much easier to rationalize these tests as flexibility tests, however, than to rationalize Match Problems as a planning test. This suggests that the Air Force planning factor was indeed *adaptive flexibility*. Another possibility is that the two factors have telescoped into one, but we prefer the interpretation stated first.

Adaptive flexibility was defined in the creativity study (13) as the ability to change set to meet new requirements imposed by changing problems. In an unpublished study in collaboration with the Institute of Personality Assessment and Research at Berkeley there was evidence of a further generality of *adaptive flexibility* in the significant loadings of two tests that required perceptual analysis and perceptual reorganization prior to solution. In that study it was also indicated that problem solving requiring insights may in general depend on the ability to shift set.

In all the tests in the above list some type of visual restructuring in order to solve a problem seems to be a common feature. Not only is the examinee required to do visual restructuring but he must also reorganize perceptually the same material in different ways to meet different requirements. In the Match Problems test, some of the problems are impossible of solution unless the examinee breaks a self-imposed instructional set. In Planning Air Maneuvers, frequently the same or similar letter configurations are used but the differences in the start and end positions among these configurations require the examinee to find new patterns.

Factor L. Ordering (Or)

21 Picture Arrangement	.53
22 Sentence Order	.45
25 Word Matrices	.44
20 Temporal Ordering	.36
31 Seeing Deficiencies	.36 (.46 J)
1 Matrix Order	.33
28 Procedure Applications	.32
41 Inference Test	.30 (.32 V, .30 LE)

This is the hypothesized *ordering* factor. Tests 21, 22, 25, 20, and 1 all require some indication of the correct arrangement of objects or events. The three tests constructed or adapted to measure the temporal aspect of ordering, Picture Arrangement, Sentence Order, and Temporal Ordering, all had significant loadings. Of the three tests that called for the hierarchical arrangement of subject matter, Outlining I and II and Word Matrices, only Word Matrices was weighted significantly.

In the two leading tests, Picture Arrangement and Sentence Order, a temporal arrangement of events is required in order to make a sensible story. In Picture Arrangement, the events are presented pictorially; in Sentence Order, the events are in language form. These two tests were involved in Adkins' reasoning study (1). In order to explain their significant loadings on a *verbal* factor in that study, it was suggested that *verbal comprehension* would be more suitably named *verbal relations* where "perception and manipulation of verbal relations" is emphasized rather than just the knowledge of word meanings. The present analysis, with the increased number of ordering tests, has effected a separation between the *verbal-comprehension* factor, defined by vocabulary tests, and *ordering*, defined by tasks in which manipulation of relations is required.

Consideration of the tests listed above leads to the interpretation of *ordering* as the ability to arrange objects or events,

or to define an arrangement of objects or events, in a sequence that is meaningful either in terms of time, hierarchical, or causal relationship.

Factor M. Elaboration (E)

7 Effects	.47 (.46 CF)
14 Planning Elaboration	.44 (.30 J)
13 Planning Skills II	.40 (.31 CF)
15 Figure Production	.37 (.30 O)
26 Unusual Methods	.34 (.32 O)

This is the elaboration factor hypothesized to be involved in tasks requiring the specification of details that would contribute to the development of an idea or to the variations of an idea.

In both Planning Elaboration and Planning Skills II, the verbal specification of details was called for in the development of plans or activities. The significant loading of Figure Production on this factor indicates that *elaboration* seems to be common to both figural and verbal tasks. In the Figure Production test, the examinee is required to produce a meaningful figure out of an ambiguous marking. The figures produced were scored for the degree of elaborateness; that is, the more detail the examinee puts in, the higher his score.

The significant loadings of the Effects test on this factor, a test in which the examinee lists four future effects of a given present trend, and of the Unusual Methods test, which calls for different ways of handling given problems, indicate that idea development can be carried on in terms of variations of an idea that show its range or scope.

Factor N. Conceptual Foresight (CF)

4 Pertinent Questions	.58
7 Effects	.46 (.47 E)
16 Alternate Methods	.44
11 Contingencies	.39 (.43 J)
8 Consequences (Remoteness)	.30 (.34 O)

The tests with significant loadings on

this factor seem to require the examinee to envision a problem situation in such ways that needs or consequences are anticipated. In tests 4 and 11 the examinee must be aware of or anticipate the needs in a situation. In tests 7 and 8 he must see the more remote consequences connected with hypothetical happenings.

The leading test on factor N, Pertinent Questions, was constructed as an orientation task in which the examinee must, by asking many different questions, show an awareness of the variables present in the situation. This was a decision-making task in which very little information was provided. The task proved to be more than an awareness process; the examinee apparently conceptualized the variable in terms of the needs he foresaw in the situation.

The Effects test was devised as a test of extrapolation, and, along with the Consequences test, was scored for remoteness. These two tests require the examinee to think of various features of the situation in terms of which the effects or consequences could be indicated. This is a task involving foresight.

Factors O, P, and Q were not defined by any loadings greater than .30, and were considered residual factors.

VI. DISCUSSION

The formulation of the hypotheses of the abilities necessary to planning performances was made in terms of the classes of planning activities and in terms of the qualitative aspects of planning. It was felt that planning could not be adequately investigated as a unitary psychological function but must be broadly conceived as a term covering a number of classes of activities and entailing a number of separate abilities. It was hypothesized that as planning proceeds through

certain stages or when different aspects of planning are emphasized, different abilities are called upon, either consecutively or in combination.

A. RELATION OF OBTAINED FACTORS TO INITIAL HYPOTHESES

The hypothesis of orientation as an activity necessary not only to planning performance but to problem-solving in general did not correspond to any one of the previously found factors or to any new ability. Orientation remains only a descriptive term for those activities that involve the seeing of an order or trend in a mass of information and that show an awareness of the pertinent variables operating in a situation. Tests of the two subhypotheses of orientation are loaded variously on factors of *ordering*, *originality*, *eduction of conceptual relations*, and *conceptual foresight*.

The futuristic character of planning performance was hypothesized to call upon a predictive ability. Tests of prediction were constructed in terms of two variations of prediction: extrapolation and foresight. For each of these variations, both perceptual and conceptual tests were included in the analysis. A separation occurred in terms of this distinction rather than between extrapolation and foresight. Competitive Planning, Symbol Grouping, and Route Planning, designed as perceptual tests involving foresight, emerged on *perceptual foresight*.

The finding of a *perceptual-foresight* factor, which is defined in part by maze tests, is in accordance with C. D. Porteus' belief that the maze tests he used were tasks involving foresight (12). Porteus implied that there is more than one level of foresight or planning, i.e., that there is a low-level kind, as typified by the ability called upon in solving the mazes,

and a more abstract or high-level kind, as called for in creative work, and that it is necessary to be capable of the lower-level foresight before one can exhibit the higher-level type. The results of this study would seem to confirm the existence of more than one kind of foresight. *Perceptual foresight* and *conceptual foresight*, however, were found as distinct factors, indicating that one type of foresight is not necessarily basic to the other.

The three conceptual tests constructed to measure the predictive ability helped to define the *conceptual-foresight* factor. Foresight in this study has been viewed as an awareness of possible future events that have a relation to a present situation. This definition of foresight appears to involve the flexibility of ideas, as indicated by the fact that the examinee has to conceptualize different needs that might be connected with the given situation in order to answer the problem in the foresight tests.

The results do not seem to support the subhypothesis of extrapolation as one type of task involving a predictive ability. The Series test, intended as a perceptual task requiring extrapolation, was not significantly weighted on any of the factors in this analysis. As indicated in the interpretation of the *conceptual-foresight* factor, the Effects test and Consequences could be characterized more as tests of foresight than of extrapolation because of the additional information required in order to give suitable responses.

Due to the multiple-response characteristic of most of the *conceptual-foresight* tests, it is not clear to what extent response fluency enters into the factor. It remains for future study to clarify this aspect of the factor.

The identification of a factor of *elab-*

oration was based upon the significant loadings of the tests connected with the subhypothesis of specification. It is an interesting finding that *elaboration* is not confined to verbal tests, as is indicated by the presence of a figural test on this factor. Apparently the ability to elaborate in a figural task is related to the ability to specify details in a verbal task. Had there been more alternate tests requiring perceptual elaboration, however, a factor for such tests might have been split off from one for conceptual elaboration.

The significant loading of the Effects test on this factor indicated that the factor should be described not only in terms of specifying details in order to develop an idea but also in terms of the production of variations of an idea which serve to show the scope of the idea.

Some comment should be made about the subhypothesis of symbolization. The elaboration factor apparently does not encompass the symbolic representation of ideas; neither does the symbolization subhypothesis define a separate symbolization factor. The expected high correlation between the two tests, Symbol Production and Line Drawing, did not materialize. The variances of these two tests were accounted for principally by the *originality* factor.

The confirmation of an *ordering* factor is indicated by the significant loadings for tests that contained items requiring the ordering of bits of information in temporal, causal, and logical sequences.

The tests designed to measure thinking ingenuity in relation to planning tasks did not serve to determine a factor of ingenuity. Much of the variance on these tests, particularly that of Unusual Methods, was accounted for by the

originality factor. There was a question whether *originality*, defined in terms of unconventionality of thinking in the creativity study, would be important in planning tasks where new procedures, or applications of old procedures to new situations, must be devised. Admittedly the hypothesis of ingenuity was not satisfactorily tested in some respects. There were too few tests representing this hypothesis, and it might have been better if the experiential features in these tests had been reduced.

Two of the three tests designed to measure the evaluation aspects of planning emerged on the *judgment* factor.

B. RELATION OF RESULTS TO FACTORS IN PREVIOUS STUDIES

Twenty tests were included in the planning battery for fourteen reference factors. Only three of the reference factors, *associational fluency*, *symbol manipulation*, and *integration III*, did not emerge in some form in the analysis. Tests of the *planning* factor found in the AAF analyses, and of *adaptive flexibility*, isolated in the creativity study, emerged together on a single factor (*adaptive flexibility*). This result helps to clear up to some degree the question regarding the essential ability represented by the AAF *planning* factor. It also broadens the conception of *adaptive flexibility*.

The results of this study were also informative with regard to the natures and scopes of certain other reference factors. *Judgment* was found to be weighted in various completion-type tests as well as tests with a multiple-choice format. The factor of *ideational fluency* seems to apply better to tasks in which responses are produced under fairly unrestricted conditions and involve responses that

arise from the more superficial aspects of situations.

A new test, Symbol Production, seems to provide a good measure of the *originality* factor, previously isolated in connection with the creative-thinking battery, and indicates a possible symbolic aspect of *originality*.

Elaboration, *ordering*, *perceptual foresight*, and *conceptual foresight* are regarded as new factors, but we should consider their possible relations to previously known factors. The verification of these factors, of course, will depend on future research. *Elaboration* may represent a new and more significant dimension of *ideational fluency*. The factor interpreted as *perceptual foresight* might alternatively be identified as the *integration III* factor found in the AAF analyses, since several of the tests that helped to define the latter factor were loaded significantly on factor J of this study. However, the leading test on *integration III*, Planning Air Maneuvers, was included in this study but did not have a significant loading on factor J. In view of this and of the fact that three tests designed to measure foresight had significant loadings on factor J, it was felt that this factor would be better characterized as *perceptual foresight*.

Conceptual foresight bears some resemblance to the *spontaneous-flexibility* factor identified in the creative-abilities study, but the idea flexibility involved in *conceptual foresight* is more purposive, or goal-directed.

C. RELATION OF RESULTS TO THE GENERAL PLANNING TEST

The Lorge test of Planning Skills (test 32) was included in the analysis as a general planning task. Given a detailed account of a practical problem situation

the examinee had to define and structure the problem for himself, find the critical areas, perhaps evaluate the alternate methods he thought of, and write down all the steps necessary for carrying out the procedures adopted. Because of the complexity of the task a number of abilities were considered to be involved.

The Planning Skills test was treated as any other variable in the analysis and no special attention was given it in the rotations of the axes. Results from the preliminary testing with two problems (only one problem situation was used in the final battery) had yielded an alternate-forms estimate of about .40. If this is a reasonable estimate, practically all of the true variance was accounted for in terms of common factors by this study. Since the factor loadings are all very low, however, our conclusions as to factor content are merely suggestive; certainly not decisive.

As expected, the common-factor variance represents a number of factors. The leading factors and the proportions of common-factor variance that each accounts for are as follows (these proportions add up to 1.00):

Factor M.	Elaboration	.22
Factor I.	Originality	.15
Factor G.	Judgment	.14
Factor F.	Ideational Fluency	.12
Factor P.	(Residual)	.18
	(Others)	.19

The Planning Skills test called for considerable detail in the examinee's report of a plan, hence the *elaboration* variance. Responses were scored to some extent for quality, therefore the production of clever and uncommon ideas would be likely to receive weight and thus introduce variance in *originality*. The contribution of the factor *judgment* can be rationalized by supposing that the examinee had to select critically the significant

features of the problem, to evaluate the hypotheses and ideas that came up, and to maintain continual check on the internal consistency of his plan. The weight in *ideational fluency* probably reflects the fact that quantity of ideas was weighted in the scoring. As indicated before, these small contributions may not be very accurately estimated. It should also be remembered that the Lorge test represents only one kind of planning.

VII. SUMMARY

The purpose of this study was to isolate and define the abilities involved in planning. A battery of 52 tests covering six hypotheses and a number of reference factors was administered to 364 USAF aircrew trainees. Included in the battery was one test that represented a general planning activity. The scores were intercorrelated and 17 factors were extracted. Orthogonal rotations resulted in 14 identifiable factors, ten of which had been found previously: *verbal comprehension*, *numerical facility*, *visualization*, *general reasoning*, *logical evaluation*, *ideational fluency*, *education of conceptual relations*, *judgment*, *originality*, and *adaptive flexibility*. Four new factors were defined as *ordering*, *elaboration*, *perceptual foresight*, and *conceptual foresight*. The greatest proportions of the common-factor variance of the general planning test were contributed by the factors of *elaboration*, *originality*, *judgment*, and *ideational fluency*.

It may be said that this investigation supports the hypothesis that in planning we should find a large number of primary abilities involved. Of the factors found in this investigation, four are new and seem to be unique to planning tests. They are the abilities of *ordering*, *elaboration*, *perceptual foresight*, and *conceptual foresight*.

REFERENCES

1. ADKINS, D. C., & LYERLY, S. B. Factor analysis of reasoning tests. Chapel Hill, N.C.: Dept. of Psychol., Univer. of North Carolina, 1951.
2. BERGER, R. M. The isolation and definition of the abilities involved in planning. Ph.D. dissertation, Univer. of Southern Calif., 1955.
3. GREEN, R. F., GUILFORD, J. P., CHRISTENSEN, P. R., & COMPREY, A. L. A factor-analytic study of reasoning abilities. *Psychometrika*, 1953, 18, 135-160.
4. GUILFORD, J. P., BERGER, R. M., & CHRISTENSEN, P. R. A factor-analytic study of planning. I. Hypotheses and description of tests. *Rep. psychol. Lab.*, No. 10. Los Angeles: Univer. of Southern Calif., 1954.
5. GUILFORD, J. P., BERGER, R. M., & CHRISTENSEN, P. R. A factor-analytic study of planning. II. Administration of tests and analysis of results. *Rep. psychol. Lab.*, No. 12. Los Angeles: Univer. of Southern Calif., 1955.
6. GUILFORD, J. P., CHRISTENSEN, P. R., KETTNER, N. W., GREEN, R. F., & HERTZKA, A. F. A factor-analytic study of Navy reasoning tests with the Air Force Aircrew Classification Battery. *Educ. psychol. Measmt*, 1954, 14, 301-325.
7. GUILFORD, J. P., FRUCHTER, B., & ZIMMERMAN, W. S. Factor analysis of the Army Air Forces Sheppard Field Battery of Experimental Aptitude Tests. *Psychometrika*, 1952, 17, 45-68.
8. GUILFORD, J. P., & GUILFORD, RUTH B. A prognostic test for students in design. *J. appl. Psychol.*, 1931, 15, 335-345.
9. GUILFORD, J. P., & LACEY, J. I. (Eds.). Printed classification tests. *Army Air Forces Aviation Psychol. Res. Rep.*, No. 5. Washington, D.C.: U.S. Government Printing Office, 1947.
10. HERTZKA, A. F., GUILFORD, J. P., CHRISTENSEN, P. R., & BERGER, R. M. A factor-analytic study of evaluative abilities. *Educ. psychol. Measmt*, 1954, 14, 581-597.
11. LUCAS, C. M., & FRENCH, J. W. The factorial composition of the relative movement test. *U.S. Navy Personnel Res. Rep.* Princeton: Educational Testing Service, July 1953.
12. PORTEUS, S. D. *The Porteus Maze Test and intelligence*. Palo Alto: Pacific Books, 1950.
13. WILSON, R. C., GUILFORD, J. P., CHRISTENSEN, P. R., & LEWIS, D. J. A factor-analytic study of creative-thinking abilities. *Psychometrika*, 1954, 19, 297-311.
14. ZIMMERMAN, W. S. A simple graphical method for orthogonal rotation of axes. *Psychometrika*, 1946, 11, 51-55.

(Accepted for publication January 25, 1957)

Infant Development Under Environmental Handicap¹

WAYNE DENNIS AND PERGROUHI NAJARIAN
Brooklyn College and the American University of Beirut

RIBBLE (10, 11) and Spitz (12, 13, 14, 15, 16) have proposed that if certain stimulus deprivations occur in early childhood the consequences are drastic and enduring. These views have arisen largely from observation of infants in institutions. The supporting evidence has consisted in part of scores of institutional subjects on infant tests and in part upon general impressions of the emotional states of the children.

This report is concerned with behavioral development in an institution whose care of infants is in some respects identical with, and in some respects quite different from, that described in other studies.

The data were obtained in a foundling home in Beirut, Lebanon, which, because of inadequate financial support, is able to provide little more than essential physical care. We will report upon the developmental status of two age-groups of children in this institution: those between 2 months and 12 months of age, and those between 4½ and 6 years of age. After describing the environmental conditions and presenting the data we will discuss the relationship of this study to previous

studies, and to theories of child development.

THE CRECHE

The institution in which the study was conducted will be called the Creche, although this is not the formal name of the home. The Creche is a home for infants and young children operated by a religious order (of nuns). All children in the Creche are received shortly after birth. They arrive via two routes. The majority come from a maternity hospital operated by the religious order referred to previously. An unmarried woman being attended by this hospital may arrange to have her infant taken to the Creche. In so doing she relinquishes claim to the infant and may not see or visit it thereafter. The remainder of the Creche population consists of infants left upon the doorstep of the institution. Nothing is known definitely concerning their parents, but it is likely that the majority of these infants, too, are illegitimate.

The Creche is nearly 30 years old but it has a new building which was completed in the spring of 1955, and for which the order is still indebted. The building is an excellent one, being fireproof, sunny, and airy. The infant beds and other pieces of equipment are new and modern. The appearance of the institution fails to reveal that it exists month after month upon inadequate and uncertain contributions. The feeding, clothing, and housing of the children have the first claim upon the Creche's meager income. The most stringent economy must be exercised in regard to expenditures for personnel. For this reason the number of persons taking care of the children

¹ This research was conducted at the American University of Beirut. The cost of the investigation was defrayed by a grant from the Rockefeller Brothers' Fund to the American University of Beirut. The writers wish to express their deep appreciation for the complete cooperation and assistance given by the staff of the Creche and by the Outpatient Department of the A. U. B. Hospital. Miss Leila Biksmati served as research assistant.

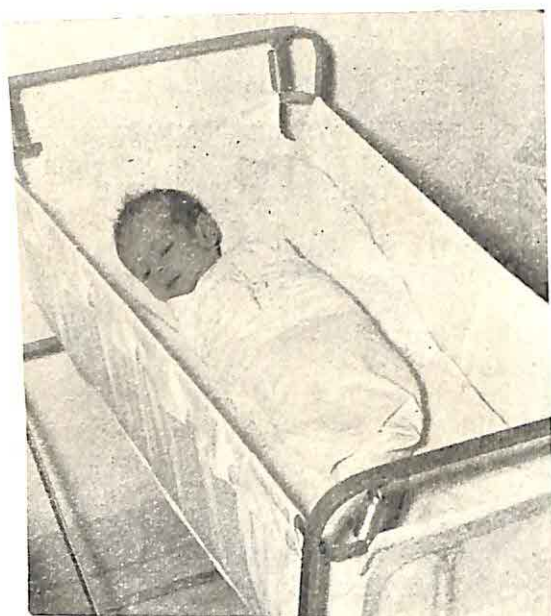


FIG. 1

is extremely limited. Understaffing is the direct cause of whatever deficiencies may characterize the child-care practices to be described later.

Naturally the number of children in the institution varies from time to time with the advent of new arrivals, and departures due to deaths, or to transfer to other institutions to which the children are sent at about six years of age. The size of the staff, too, is subject to some variations. However, estimates made at two periods separated by five months agree in showing that for each person directly concerned with the care of the children—i.e., those who feed the children, change diapers, bathe and clothe them, change their beds, nurse them when they are ill, supervise their play, and teach them—there are 10 children. This ratio of 1 to 10 includes those on night duty as well as on day duty. It does not, however, include personnel who work in the kitchen, laundry, and mending room, nor those who do the cleaning. It does not include the four nuns who constitute the administrative staff and who frequently assist in direct care. Clearly this is an extremely limited staff. The essential functions can be accomplished only by means of hurried procedures and long hours of work.

From birth to one year there is no assignment of individual children to particular attendants. Rather, a room of children is assigned jointly to several caretakers and observation showed no consistent relationships between attendants and children. At later ages, each group of children is assigned most of the day to a supervisor and an assistant.

During the first two months of life the infant

is taken out of his crib only for his daily bath and change of clothes. He is given his bottle while lying on his back in his crib, because ordinarily no one has time to hold it. The nipple is placed in his mouth and the bottle is propped up by a small pillow. Bathing and dressing are done with a maximum of dispatch and a minimum of mothering.

In conformity with a widespread Near Eastern practice, the infant is swaddled from birth. Figure 1 illustrates the type of swaddling used. The baby has his arms as well as his legs enclosed in tight wrappings, and hence the scope of his movements is greatly restricted. During the early weeks the infant is bound as depicted except when being bathed and dressed. No fixed schedule is followed in regard to freedom from swaddling, but in general the hands are freed at about two months of age, and swaddling is ended at about four months. Swaddling is continued for a longer period during the winter months than during the remainder of the year because the wrappings of the child serve to keep him warm.

As shown in Fig. 2, each crib has a covering around the sides. This is present to protect the child from drafts, but as a consequence the child can see only the ceiling and the adults who occasionally come near him.

The adults seldom approach him except at feeding times. When they feed him they do not usually speak to him or caress him. When two or three persons are feeding twenty infants, many of them crying, there is no tendency to dally.

At about four months of age the child is removed to a room for older infants. He is placed in a larger crib, but for several further months his care remains much the same as it has been.

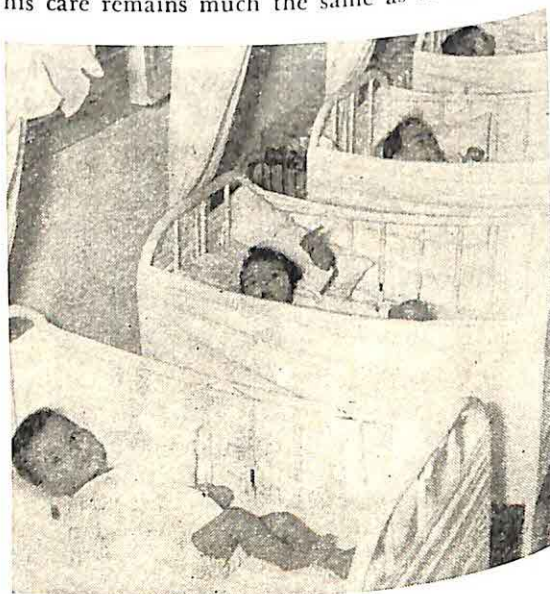


FIG. 2

A typical scene in the room is shown in Fig. 2. A toy is usually placed in each crib, but it soon becomes lodged in a place inaccessible to the child and remains there. The child remains in this second crib until he begins to pull to the edge of the crib and faces some danger of falling out. At this point, he is usually placed during his waking hours with one or two other children in a play pen. This situation is illustrated in Fig. 3. Sometimes he is placed in a canvas-bottomed baby chair, as shown in Fig. 4, but this is usually done only for short periods of time. The older child takes his daytime naps in the play pens. He is returned to his crib at night and tightly tucked in. The child graduates from room two to another room at one year of age or slightly thereafter. Some description of the care of older children will be given on later pages.

Until about four months of age the infant's food consists of milk, supplemented by vitamins. The feedings during this time are on a schedule of six feedings per day at daytime intervals of three hours. After four months bottle feeding is gradually reduced in frequency. It ceases at about twelve months.

The introduction of cooked cereals begins at four months, and fruit juices, crushed bananas, apple sauce, and vegetables are begun at five months. Depending upon the preferences of an attendant a child is sometimes given these supplementary foods held in arms, sometimes while sitting in chairs, and sometimes lying down. Beginning at eight months, eggs and chopped meat are occasionally given. Feeding times are reduced to five times per day at four months and to four times per day at one year. Toilet training is begun between 10 and 12 months.

Children are weighed at weekly intervals. Seri-



FIG. 3



FIG. 4

ous efforts are made to give special feeding to infants who are not gaining properly but again staff limitations make it difficult for an attendant to spend much time with any one child. The average weight during the first six months, based on records of the infants which we tested, is appreciably below what is ordinarily considered desirable (see Table 1). Comparable data are not available for other Lebanese children. No data are available on children beyond six months of age at the Creche.

From about one year to about three years the children spend much of the day in play groups of about twenty children with a supervisor and an assistant. Equipment is limited to a few balls, wagons, and swings. From three to four years of age much of the day is spent seated at small tables. The children are occupied in a desultory way with slates, beads, and sewing boards. At about four years they are placed in kindergarten within the Creche where training in naming objects and pictures, writing, reading, and numbers is begun. Instruction is given in both Arabic and French.

Diet and medical care are under the supervision of a physician who devotes, gratis, about one hour per day to the Creche, whose population is about 140 children. During the winter months colds are common, and pneumonia occasionally occurs. The usual childhood illnesses occur. When a contagious disease enters the Creche it is likely to become widespread since there are no facilities for isolation of infectious cases. We do not have adequate statistics on mortality. It is our impression that it is high in the first three months of life, but not particularly high thereafter. Mortality seems especially high among those infants who are found on the doorstep, many of whom are suffering from malnu-

TABLE I
AVERAGE WEIGHTS OF CRECHE INFANTS

Statistic	Boys						
	Birth	1 mo. ^a	2 mo.	3 mo.	4 mo.	5 mo.	6 mo.
Average weight in grams	2926	3233	3746	4365	4926	5555	5984
Number of cases	28	28	28	27	23	18	16
	Girls						
	Birth	1 mo.	2 mo.	3 mo.	4 mo.	5 mo.	6 mo.
Average weight in grams	2727	2985	3353	3858	4436	4910	5463
Number of cases	13	12	13	13	11	10	8

^a In computing this average, for each child the record of weight taken nearest age 1.0 month was employed. A similar procedure was used at other ages.

trition, exposure, or disease upon admission. In evaluating institutional mortality it should be noted that in some areas of Lebanon the crude death rate in the first year among children in homes is as high as 375 per 1000 (6).

THE COMPARISON GROUP

For comparison with behavioral records of the Creche infants, data were obtained from children brought to the Well Baby Clinic of the American University of Beirut Hospital. All well babies of appropriate age who were brought to the clinic on certain days were tested. They were from among the poorer, but not the poorest, segments of the Beirut population.

All children tested were living at home and were brought to the clinic by their mothers. The majority were being breast fed. We did not obtain detailed data on swaddling, but typically the younger babies were brought in swaddled and the older ones unwaddled. It is our impression that swaddling customs among the poorer half of the Beirut population approximate those of the Creche. This conclusion is supported by a study by Wakim (17). Other comparison data were provided by American norms and certain Lebanese norms to be described later.

THE TESTING PROGRAM

For the subjects under one year of age the Cattell infant scale was employed (2). This scale was selected because among available tests it seemed to offer

the most objective procedures for administration and scoring. It provides five items for each month from two to 12 months of age, with one or two alternate items at each age level.

The procedures described in the test manual were carefully followed. They call for testing each infant at a level at which he passes all tests, at a level he fails all tests, and at all intermediate levels.

Several items on the test were not applicable to the Creche group because they require the examiner to obtain information from the mother or other caretaker. Among such items are babbles, anticipates feeding, inspects fingers, says "dada," etc. Attendants at the Creche could not supply the information required by these items. For this reason, "alternate" items provided by Cattell and based on direct observation were regularly substituted for these items. In the case of the comparison infants, all age-appropriate items, including all alternates, were administered; but in computing developmental scores for comparative purposes identical items were used for the Creche and the comparison groups. At the 4½-to-6-year level the tests used were the Goodenough draw-a-man test,

the Knox cube test, and the Porteus maze test. These were chosen because it was judged that they might be but little affected by the environmental handicaps of the Creche children. They have the further advantage of requiring a minimum of verbal instructions.

In giving and scoring the draw-a-man test, Goodenough procedures (5) were followed. For the other two tests the procedures and norms employed were those given in the Grace Arthur Scale of Performance Tests, Revised Form II (1).

NUMBER OF SUBJECTS

We tested all subjects who fell into our age categories upon two series of testing dates. The only exceptions consisted of children who were ill or who had just undergone serious illness. The infant tests were given to 49 Creche infants and the 41 comparison cases. Since rather few of the Creche infants were above six months of age at the time of our first period of testing, during our second testing period we tested all infants who were six months of age and over even though this meant retesting in 13 cases. For this reason the number of test scores for the 49 Creche infants is 62.

In the 4½-to-6-year group, Goodenough tests were given to 30 subjects, and the Knox cube test and the Porteus maze test were each given to 25 subjects. None was retested.

RESULTS

For the infants, Table 2 indicates by age levels the score earned on each test. The Creche scores are shown by O-symbols, the comparison scores by X-symbols. Scores are grouped by step intervals of ten points. Thus, examining the figure by beginning at the top of column one, one finds that between 2.0 and 2.9 months of age one comparison infant had a develop-

mental quotient between 140 and 149, two comparison infants had quotients between 130 and 139, etc.

Examination of Table 2 shows that at the two-months age level there is little if any difference between the two groups. The mean of the Creche group is 97, that of the comparison group 107. These means, each based on only 8 cases, are not significantly different from each other or from the American norms. However, at all ages beyond 3.0 months the Creche infants score definitely lower than either the comparison or the normative groups, whose records are indistinguishable.

If all scores from 2 to 12 months are averaged, the Creche mean is 68, the comparison mean 102. For the 3-to-12-month period the mean of the Creche scores is 63, (*SD* 13), that of the comparison group 101 (*SD* 15), a difference of 38 points. This is a very large and highly significant difference ($P < .001$). In this age range all of the comparison infants tested above the mean of the Creche subjects and all of the Creche subjects were below the mean of the comparison group. No Creche baby between 3 and 12 months had a DQ above 95.

Before discussing the results of the infant tests we turn now to the tests given to Creche children between 4.5 and 6 years of age. We note first that there are reasons to believe that the subjects tested at 4.5 to 6.0 years of age performed, as infants, at the same level as did the children whose test results have just been presented. Because procedures of admission to the Creche have not changed in recent years the two groups of infants can be assumed to be genetically similar. Since practically all infants who enter the Creche remain for six years, there are no selective influences between admission and six years. The only qualification of this statement regards infant mortality,

TABLE 2
INDIVIDUAL INFANT SCORES BY AGE^a

Scores	Age in Months									
	2	3	4	5	6	7	8	9	10	11
140-149	X									
130-139	XX		X							
120-129	X	X					X			
110-119	00		X		XX	XX		X		
100-109	00XX	XXXX	X	XX	X	XX	X		XX	
90-99	00	O		XX	OX					
80-89	00X	OX	XXXXX		O			O		O
70-79		00	00X		00			O	00	
60-69		O	000000	000000	O	X	O	O	O	O
50-59	X	00	000	00	O	O			00	00
40-49			O		00	000		00	O	

^a Creche infant scores are indicated by O; comparison infants by X.

whose selective action so far as psychological tests are concerned is unknown, here as elsewhere. According to the supervisory staff there have been no changes in child care within the past six years.

The results of the performance tests are shown in Table 3. It will be noted that the data there reported agree remarkably well in showing that on these tests the development of the Creche children is only about 10 per cent below the norms of American home-reared children. In a separate report (3) it has been shown that on the Goodenough test Lebanese children at the five-year level make scores equivalent to the American norms. No Lebanese norms are available for the

Knox cube or Porteus maze tests but there is no reason to believe that they would be higher than the published standards. In other words, there is evidence that the environment of the Creche produces only a slight retardation among four- and five-year-olds on these tests.

In summary, the data show that, with respect to behavioral development, children in the Creche are normal during the second month of age, are greatly retarded from 3 to 12 months of age, and almost normal on certain performance tests between 4.5 and 6 years of age.

INTERPRETATIVE DISCUSSION

To a reader acquainted with the numerous and often divergent opinions concerning the effects of early environment, the results just reported may, on the surface, only serve to confuse further the already unclear picture. We believe, however, that we can show that these data and others can be fitted into a coherent view.

Early Normality of Creche Infants

The fact that the Creche subjects had DQ's of approximately 100 during the second month, and presumably during the first month also, should not be sur-

TABLE 3
RESULTS OF PERFORMANCE TESTS

Test	Various "DQ" Scores			
	N	Range	Median	Mean
Goodenough	30	58-136	93	93
Porteus maze ^a	25	69-150	89	95
Knox cube ^b	25		100	

^a Four children earned fewer than 4 points, which is the minimum score for which Arthur gives a mental age. Since the lowest MA given by Arthur is 4.5, these children were arbitrarily given a mental age of 4 years and DQ's were computed accordingly. Obviously these scores affect the mean and SD but not the median.

^b On this test, 11 of the 25 subjects scored below the 4.5 MA, the lowest age for which Arthur gives norms. Because of the large number below 4.5 no arbitrary scores were given. Of the 14 subjects who earned MA's of 4.5 and above, one had a DQ of 80 and two of 100. The remaining scores ranged from 101 to 165. The median of 100 seems representative.

prising. It has not been shown that any stimulus deprivation will affect infant behavioral development during the first two months. The twins reared under experimental conditions by Dennis and Dennis (4) made normal progress during this period. The infants tested by Spitz (12) had a mean developmental quotient of 130 during the second month. The supernormality of this score was probably due to the inadequacy of test norms rather than to institutional influence.

If it is true that restricted stimulation has little or no effect upon early behavioral development, this can be due to at least two different causes. One explanation would be in terms of maturation. Perhaps growth of the nervous system, apart from sensory stimulation, is alone responsible for postnatal behavioral growth during the first two months. A second explanation lies in the possibility that sensory experience is essential, but that for the tests presented to him the infant even when swaddled hand and foot and lying on his back obtains sufficient stimulation.

For the Cattell infant tests the second interpretation is not altogether unreasonable. Of the five tests which we employed at the two-months level, four are given to the infant while lying on his back and the responses required are visual. These are "inspects environment," "follows moving person," "follows moving ring vertically," and "follows moving ring horizontally." Since the infants spend nearly 24 hours per day in a supine position in a well lighted room, and some movement occurs near them, there is considerable opportunity to practice visual pursuit movements.

The fifth item among the two-month tests is lifting head when prone. The Creche infants are placed on the abdomen for a short time daily while being bathed, dried, and dressed. For this reason, lifting the head while in this position can be practiced and direct observation shows that it is practiced. Possibly the Creche infants respond normally to the items given them at two months because the required responses are well practiced. However, the possibility that maturation alone is sufficient for the development of the items is not ruled out.

Retardation Between Three and Twelve Months of Age

Beyond the two-months level the majority of items on the Cattell scale require that the infant be tested in a sitting position while being held on the lap of an adult. Sitting is a position to which the Creche infants under about ten months of age are relatively unaccustomed. They are not propped up in their beds or placed in chairs before that age. The first occasion for placing the infants in a sitting position may come with the introduction of semisolid foods, but we have noted that some of the infants are given these while lying down. Perhaps as a consequence of inexperience in being held upright the infants as a group make a poor record on the test item which involves holding the head erect and steady. This unsteadiness of the head, plus general unfamiliarity with sitting, may account in part for the low scores earned on certain purely visual items. These are "regards cube," "regards spoon," "follows ball," and "regards pellet."

Many of the remaining items involve not only sitting but in addition manual skills directed by vision. Among the items are "picks up spoon," "picks up cube," "grasps pellet," "grasps string," "lifts cup," "takes two cubes," "exploits paper," "pulls out peg," etc. Between ages five and seven months, the age placement given these items, the infants have little opportunity to practice visuo-manual coordinations in a sitting position and, further, visuo-manual coordinations are not required or encouraged even in a lying position.

Analysis of other items whose placement is between three and twelve months reveals that practically all of them require manual skills and require adjustment to visually presented objects. It is suggested that the relationship between

the items and the environmental restrictions experienced by the children account for the low scores made by the Creche subjects.

We examined the records made by the Creche children aged 3 months and above on each item, expecting that one or two items might be found in regard to which their performance is normal. We were able to find none. But we were also unable to find an item in this age range on which the subjects were judged to receive a normal amount of relevant experience.

It is interesting to note two items on which the subjects are very deficient even though the motor component of the item is clearly present. These involve turning to sound. In one of these items, the child, sitting on the lap of an adult, is required to turn toward the experimenter who stands by the shoulder of the seated adult, and calls the infant's name. The second item is similar but a small hand-bell is used instead of the voice. The first item has an age placement of four months, the second, five months. Of 36 children tested between 4.0 and 10.0 months of age only one turned to the voice and only 4 turned to the bell.

Now all of the children turned to and followed a moving person in the field of view. The difficulty of the item apparently lies in the subject's lack of associations with sounds. We have noted that in approaching a child or providing services for a child the attendants seldom speak to him. This seems to be due partly to the fact that the attendants are too busy. A second relevant fact is that, with 20 children in a room, and the windows open to rooms containing 100 additional children, it is seldom quiet enough at feeding times and bathing times to encourage verbal greetings. So far as we could determine no event which happens to a Creche baby is consistently preceded by a sound signal. These conditions seem to explain the finding that the infants seldom turned to a voice or a ringing bell only a few inches from their ears.

From the preceding discussion it will be obvious that we tend to attribute the retardation of Creche subjects between 3 and 12 months of age to a lack of learning opportunities relative to the Cattell test items.

Relationship of the 3-to-12-Month Retardation to the Findings of Other Studies

There seems to be a superficial, if not a basic, disagreement between the results here reported, and those of other studies, particularly those of Dennis and Dennis (4) and those of Spitz (12-16). We wish to comment on the apparent divergences and to indicate how they can be reconciled.

In a study of a pair of twins named Del and Rey who were reared under experimentally controlled conditions until thirteen months of age, Dennis and Dennis found that, while the subjects were retarded beyond the range of ordinary subjects in regard to the appearance of a few responses, the subjects' development in general equalled that of home-reared infants. The few specific retardations occurred on items in respect to which the infants could not engage in self-directed practice, namely, visually directed reaching, sitting without support, and supporting self with the feet. These retardations seem consonant with the behavior of the Creche subjects. However, the prevailing normality of Del and Rey seems at variance with the Creche findings.

To begin with, certain differences between the environmental conditions of the subjects in the two studies should be noted. For one thing, the adult-child ratios in the two studies were very different. In the Del-Rey study there were two subjects and two experimenters, a one-to-one ratio. In the Creche, the adult-child ratio is one to ten, a greatly different situation. In the Dey-Rey study the environmental restrictions in regard to learning were rather severe in the beginning, but were gradually relaxed as desired data were obtained. In the Creche, very limited opportunities for learning

and practicing responses continue throughout the first year. Certain specific contrasts may be mentioned: Del and Rey were kept in larger and deeper cribs, were less restrained by clothing and consequently probably had more opportunities for motor experimentation than did the Creche infants. Further, Del and Rey may have received more handling and more varied exposure to stimuli than did the Creche infants. However, there can be no doubt that in several respects Del and Rey suffered as much a restriction of experience as did the Creche infants. Speech was not directed to Del and Rey nor did adults smile in their presence until they were six months of age. No toys were provided until the twelfth month. They were not placed in a sitting position until they were over eight months of age.

But it is our belief that the difference between the normality of Del and Rey and the retardation of the Creche infants is due to the use of different indices of behavioral development rather than to real differences in behavior. In the Del-Rey study no general scale of infant development was administered. The majority of the developmental data reported for Del and Rey consisted of noting when each of a number of common infant responses first appeared. That is, the observers recorded when each subject first brought hand to mouth, first grasped bed-clothes, first vocalized to person, first laughed, etc. The initial date of occurrence of such responses cannot be determined by testing. The Del-Rey data are longitudinal and the Del-Rey records were found to be normal when compared with similar data obtained in other observational studies.

Now since observation in the Del-Rey study was directed primarily toward responses which could occur at any time

and did not require the introduction of test conditions, it follows that poverty of environmental stimulation would not be expected to yield much evidence of retardation. The child left to his own devices on his back in his crib can bring his hand to his mouth, grasp his bed clothes, vocalize, observe his own hands, grasp his own hands, grasp his own foot, bring foot to mouth, etc. These are the items which were observed. One of the major findings of the Del-Rey study was that the untutored infant does do these things, and does them within the usual age range of home-reared babies.

In regard to such responses it *may* be that the Creche babies are normal. The relevant facts can be discovered only by observers each spending full time observing a few infants. If all Creche infants were to be observed it would necessitate the presence of many additional observers or caretakers. The reader is reminded that the Del-Rey investigation, involving only two infants, took a major part of the time of the two observers for one year. To devote one year, or even one month to observing each Creche subject cannot be proposed in an institution which has severe limitations of caretaker personnel. In contrast to the requirements of an observational study of development, the testing time in the Creche study was only 10 to 30 minutes per subject.

If we cannot compare Del and Rey with the Creche babies in terms of observational data, it is likewise not possible to compare them in terms of test data. It is impossible to estimate in retrospect with any degree of confidence how Del and Rey would have scored at various times during the first year on the Cattell Infant Scale. We arrive, therefore, at the following conclusion: It is likely infants with restricted learning opportunities are normal on "observational" items but re-

tarded on "test" items. It is believed that the latter, but not the former, are influenced by environmental limitations. If this is a correct interpretation, the Del-Rey study and the Creche study are two sides of the same coin. However, to establish that this is the case appears to be a very difficult research assignment.

We consider next the work of Spitz. The observations by Spitz which seem most closely related to the present study concern the institution called Foundling Home. Here, as at the Creche, there was a shortage of personnel. Although the mothers were present in the institution for several months, they seem to have had little contact with their children aside from breast-feeding them. Pinneau (9) points out that Spitz does not explain why this was the case. Despite the presence of the mothers in the institution the adult-child ratio in the nursery is reported to be about 1 to 8. The children spent most of their time for many months on their backs in their cribs, as did the Creche infants. At one point Spitz reports that a hollow worn in their mattresses restrained their activity. This, however, was definitely not true of the Creche infants.

Since Spitz's studies have been extensively reviewed and criticized by Pinneau, only a limited amount of space will be devoted to them here. Spitz used some form of the Hetzer-Wolf baby tests. There is no doubt that their standardization leaves much to be desired. Spitz reports scores for the Foundling Home group and a control group of 17 home-reared infants. In the second month both groups had mean DQ's between 130 and 140. The private home group remained at that level but the mean of the Foundling Home group dropped precipitously to 76 by the sixth month and to 72 by the end of the first year. Spitz believes

that this decline in DQ was due to the emotional consequences of separation from the mother, but Pinneau has pointed out that most of the decline took place prior to the prevalent age of separation. Pinneau indicates further that at least some of the decline is probably due to inadequate test standardization.

We compare our data with those of Spitz with considerable hesitation because the two sets of data were obtained by tests whose comparability is unknown. In numerical terms the results of the two studies in the second half of the first year seem to agree fairly well, Spitz's mean for this period being about 74 and ours 63. But the findings for the first half-year present some apparent differences. Our subjects drop from a mean of 97 to a mean of 72 between the second and third months, and drop only ten additional points thereafter. Spitz's group starts higher and declines for a longer period.

Spitz's data and ours agree in finding that environmental conditions can depress infant test scores after the second month of life. We disagree with Spitz in regard to the interpretation of the cause of the decline. He believes it to have been due, in the case of his subjects, to a break of the emotional attachment to the mother. This could not have been the cause of the decline of the Creche infants. Since the conditions for the formation of an emotional tie to a specific individual were never present, no breach of attachment could have occurred. We have noted above Pinneau's demonstration that even Spitz's own data do not support his interpretation. We believe that Spitz's data as well as ours are satisfactorily interpreted in terms of restricted learning opportunities. We suggest that an analysis of the relationship between test items and the conditions prevailing in the Foundling Home would reveal

that retardation could readily be explained in terms of restriction of learning opportunities. But such restriction is not inherent in institutional care. Klackenberg has recently presented a study (7) of infant development in a Swedish institution, in which the adult-child ratio was 1 to 2 or 3, in which no retardation was found.

Discussion of the Creche Four- and Five-Year Olds

We have no doubt that on many tests the Creche four- and five-year-olds (and also two- and three-year-olds) would be retarded, perhaps to a marked degree. We think this would be particularly true in regard to tests involving more than a very modest amount of language comprehension and language usage. The language handicap of institutional children with limited adult contact has been sufficiently demonstrated (8).

It is likely that on some performance tests the Creche children also would score below available norms. On the Healy Picture Form Board, for example, most of the incidents represented are outside the experience of Creche children. We assume that the older Creche children are retarded on some tests, but we wish to determine whether retardation is general or whether it is related to specific environmental handicaps.

We chose the draw-a-man test, the Knox cube test and the Porteus maze test because it was thought that the Creche environment might affect these tests less than other tests. So far as the Knox cubes are concerned, it is difficult to imagine how one can deprive a child of the experience of visually remembering just-touched objects, except through loss of sight. So far as the Goodenough is concerned, both human beings and two-dimensional representations of them were

familiar to the subjects. They were also familiar with the idea of drawing and with the use of pencils. Knowledge of the use of pencils may also play a part in the Porteus maze test. It is uncertain what other experience may play a role in this test.

The results show clearly that on these tests the Creche children approximated the performance of children in normal environments. In other words, the retardation which was found to exist between 3 and 12 months of age did not produce a general and permanent intellectual deficit. It is possible for infants who have been retarded through limitations of experience at an early age level to perform normally, at least in some respects, at later age periods. The assumption that early retardation produces permanent retardation does not receive support from our data.

Emotional and Personality Effects

No doubt many readers would like to know the emotional and personality consequences of the Creche regime. So would we. But to the best of our knowledge no objective and standardized procedures with adequate norms are available which would enable us to compare the Creche infants with other groups of children in these respects. This is equally true of studies conducted earlier.

In the absence of objective techniques, we can only report a few impressions. The Creche infants were readily approachable and were interested in the tests. Very few testing sessions were postponed because of crying, from whatever cause. There was very little shyness or fear of strangers, perhaps because each infant saw several different adults. In the cribs there was very little if any crying that did not seem attributable to hunger or discomfort. However, some of the older babies developed automatisms such as arching the back strongly, or hitting some part of the body with the hand, which may have represented a type of "stimulation hunger." It was almost always possible to get the infants over two months of age to smile by stroking their chins or

cheeks or by shaking them slightly. The older children, like the infants, were friendly and approachable. However, such observations are not meant to imply that other personality consequences could not be found if adequate techniques existed.

SUMMARY AND CONCLUSIONS

This study has been concerned with the development of children in an institution in Beirut, Lebanon, called the Creche, in which "mothering" and all other forms of adult-child interaction are at a minimum because the institution is seriously understaffed. The children come to the institution shortly after birth and remain until six years of age. Contact with the mother ceases upon the child's entrance to the institution and contact with mother-substitutes is slight because the adult-child ratio is 1 to 10.

Opportunity for developing infant skills through practice is very slight. In the early months the infants are swaddled. For many months the infant lies on his back, and is even fed in a supine position. He is not propped up, carried about, or provided with the means of practicing many activities.

Data on behavioral development were obtained by giving the Cattell infant scale to all infants between two and twelve months of age and the Good-enough draw-a-man test, the Knox cube test, and the Porteus maze test to all children between $4\frac{1}{2}$ and 6 years of age. Comparison data were available from American norms and from certain groups of Lebanese subjects.

It was found that in terms of developmental quotients, the mean quotient at two months was approximately 100. Between three and twelve months the mean

was 63. In the tests given at the four- and five-year level, the mean scores were roughly 90.

Possible interpretations of these data have been discussed at some length. Our conclusions may be summarized as follows:

1. It is uncertain whether the normality of behavior at two months shows that maturation plays a major role in early development, or whether experience, limited as it was, provided the essential requirements for learning the responses which were tested.
2. The retardation prevailing between three and twelve months of age seems to be due to lack of learning opportunities in situations comparable to the test situations. It is possible that an observational approach in the day-by-day situation might reveal that some behaviors developed normally.
3. The infants did not undergo loss of an emotional attachment. There is nothing to suggest that emotional shock, or lack of mothering or other emotion-arousing conditions, were responsible for behavioral retardation.
4. Retardation in the last 9 months of the first year to the extent of a mean DQ of 65 does not result in a generally poor performance at $4\frac{1}{2}$ to 6 years, even when the child remains in a relatively restricted environment. The study therefore does not support the doctrine of the permanency of early environmental effects.
5. It is believed that the objective data of other studies, as well as this one, can be interpreted in terms of the effects of specific kinds of restrictions upon infant learning.

REFERENCES

1. ARTHUR, GRACE. *A point scale of performance tests* (Rev. Form II). New York: Psychol. Corp., 1947.
2. CATTELL, PSYCHE. *The measurement of intelligence of infants and young children*. New York: Psychol. Corp., 1940.
3. DENNIS, W. Performance of Near Eastern children on the draw-a-man test. *J. educ. Psychol.*, in press.
4. DENNIS, W., & DENNIS, M. G. Development under controlled environmental conditions. In W. Dennis (Ed.), *Readings in child psychology*. New York: Prentice-Hall, 1951. Pp. 104-131.
5. GOODENOUGH, FLORENCE. *The measurement of intelligence by drawings*. New York: World Book Co., 1926.
6. KHAMIS, S. H., & POWERS, L. E. Report on infant mortality survey of rural Lebanon. Mimeographed report, Amer. Univer. of Beirut, June, 1955.
7. KLACKENBERG, G. Studies in maternal deprivation in infants' homes. *Acta paediatr.*, 1956, 45, 1-12.
8. MCCARTHY, DOROTHEA. Children's speech. In Carmichael (Ed.), *Manual of child psychology* (2nd rev. ed.). New York: Wiley, 1951.
9. PINNEAU, S. R. The infantile disorders of hospitalism and anaclitic depression. *Psychol. Bull.*, 1955, 52, 429-452.
10. RIBBLE, MARGARET. *The rights of infants*. New York: Columbia Univer. Press, 1943.
11. RIBBLE, MARGARET. Infantile experience in relation to personality development. In J. McV. Hunt (Ed.), *Personality and the behavior disorders* (Vol. 2). New York: Ronald Press, 1944. Pp. 621-651.
12. SPITZ, R. A. Hospitalism. An inquiry into the genesis of psychiatric conditions in early childhood. *Psychoanal. Stud. Child*, 1. New York: Int. Univer. Press, 1945. Pp. 53-74.
13. SPITZ, R. A. Hospitalism: A follow-up report. *Psychoanal. Stud. Child*, 2. New York: Univer. Press, 1946. Pp. 113-117.
14. SPITZ, R. A. Anaclitic depression. *Psychoanal. Stud. Child*, 2. New York: Int. Univer. Press, 1946. Pp. 313-342.
15. SPITZ, R. A., & WOLF, KATHERINE M. Autoerotism. Some empirical findings and hypotheses on three of its manifestations in the first year of life. *Psychoanal. Stud. Child*, 3-4. New York: Int. Univer. Press, 1949. Pp. 85-120.
16. SPITZ, R. A. The psychogenic diseases in infancy: An attempt at their etiologic classification. *Psychoanal. Stud. Child*, 6. New York: Int. Univer. Press, 1951. Pp. 255-275.
17. WAKIM, S. Child care in Mich-Mieh. M.A. Thesis, Library of the Amer. Univer. of Beirut, June, 1956.

(Accepted for publication January 28, 1957)

Gradients of Error-Reinforcement in a Serial Perceptual-Motor Task^{1,2}

MELVIN H. MARX

University of Missouri

I. INTRODUCTION

THE primary problem of the present study was to determine whether a statistically significant gradient of response strength of errors around rewarded responses would remain after removal of the guessing-sequence and probability-bias factors by control groups directly comparable to the experimental groups.

Originally reported by Thorndike (16, 17), the "spread-of-effect" phenomenon has received substantial empirical verification. Considerable disagreement, however, has developed about its theoretical interpretation. Thorndike (16) regarded the gradient of decreasing intertrial repetition of errors with increasing distance from a rewarded response as strong evidence for the *direct* strengthening by reward of S-R connections. He saw it as an "independent proof" of his law of

effect. More recent investigators have presented a wide variety of alternative interpretations. These range from the suggestion that the Thorndike effect is simply one special case of a more inclusive generalization-of-reinforcement principle (9) to the conclusion that the obtained gradients bear no essential relation to *reward* but are artifacts of such conditions as number-guessing habits and probability bias in the response series (3, 4, 14, 15).

The present study is an attempt to provide data on this problem with more adequate controls than have hitherto been used. The necessity for such controls is pointed up by the various studies which have shown that artifactual gradients can be produced under certain conditions in the absence of reward. However, the gradients that have been obtained under conditions designed to test the guessing-sequence and the probability-bias hypotheses cannot be directly compared with those obtained under the usual conditions of reward following motivation to learn. The evidence thus far obtained may be used to support the conclusion that these alternative factors can, under certain conditions, produce gradients; but these kinds of data cannot answer the question as to whether these factors are *necessary* or *sufficient* to explain the typical Thorndike gradients obtained within different experimental frameworks. These results do, of course,

¹The data for this study were collected during the writer's period of employment as Research Psychologist at the Human Resources Research Center (now the Air Force Personnel and Training Research Center), Lackland Air Force Base, Texas. The opinions and conclusions contained herein are those of the author and are not to be construed as reflecting the views or the endorsement of the Department of the Air Force. A partial report of the study was given at the Chicago meeting of the American Psychological Association in September, 1951. (Abstracted in *Amer. Psychologist*, 1951, 6, 293.)

²I am greatly indebted to Dr. Shinkuro Iwahara, for his assistance in supervising most of the statistical work, and to W. A. Hillix, for his assistance in the preparation of the manuscript.

throw suspicion on the Thorndike interpretation and point to the need for direct, controlled comparisons of the sort attempted in the present investigation. This problem is discussed at greater length in the recent review of the spread-of-effect literature (8).

II. METHOD

A. Subjects

Four groups of 300 basic trainees at Lackland Air Force Base were used. Two of these were experimental groups, E and EU; two were control groups, C and CRB. Data for the first two trials from an additional 74 Ss were also used for certain of the analyses, since these Ss were treated exactly like those of Group C, who had been tested earlier. This gave a pooled Group C of 374 Ss and a total of 1,274 Ss in the study.

All of the Ss were basic trainees from flights at Lackland Air Force Base between October 1950 and August 1951. They were all tested individually by the same airman, who was thoroughly trained as to the testing procedure but naive as to the purpose of the study. Ss were made available from the various flights according to the assignments worked out by the airman in charge of their distribution and were then placed in the various groups according to a mechanical system of rotation which was prepared in advance.

Throughout the study it was found necessary to discard only 20 Ss. Most of these eliminations were for some misunderstanding of the instructions or experimental error in test administration. These cases were discarded by E before the records were analyzed.

B. Apparatus

A SAM complex coordination Test, Form E (11), was modified in accordance

with the requirements of this study. Only the aileron lights were used. A plywood mask screened off the elevator and rudder lights to avoid distraction of the S. The resulting display is shown in Fig. 1, which is a frontal view of the apparatus (not drawn to exact scale).

Movements of the stick from the central resting position to right or left were accompanied by the flashing on of the appropriate green lights in the lower row of lights. The middle light, representing the resting position of the stick, was not used. This left a total of 12 lights available for settings, 6 to the right and 6 to the left of the central resting position. The top row of red lights was used only as a signal. These lights were flashed on, as a group, for the "right" signal.

The normal tension on the stick was used. S was instructed to allow the stick to return to its central resting position as soon as he had made a response and had been given the right or wrong signal.

The rudder bar was fitted with a microswitch and wired into the control panel to make possible knowledge of results after each response. The S was instructed to push the rudder bar with his right foot immediately after making a light setting. Flashing on of all of the red lights in the upper row—the signal for right—or activation of a buzzer—the signal for wrong—then occurred as long as the bar was depressed.

A simple control panel was arranged for E, who sat at a table immediately behind the display panel. This consisted of a row of 12 lights corresponding to the 12 light settings available to S, and enabled E to observe S's responses as they were made. A master control switch, silent in operation, enabled E to arrange the right or the wrong signal in advance of each response.

This apparatus made it possible to use all the data. In almost all the research thus far reported on this problem, response strength has been measured only by the percentage of repetition of responses (cf. 8). This all-or-none type of measure is more difficult to treat statistically and less sensitive than a measure in which tendencies toward repetition of response can be quantitatively scaled. Such a quantitative scaling was made possible in this study by using the variation, in linear light units, between settings made on successive trials. Use of this measure

was based on the assumption that the stronger responses would be less varied from one trial to the next, with exact repetition as the limiting case.

The task was also designed to minimize systematic pre-experimental response sequences of the sort which have been observed in the number-guessing studies and which have formed the basis for the guessing-sequence explanation of the Thorndike gradients.

C. Experimental Design

1. General Testing Procedure³

Instructions for both experimental groups and for the standard control group (C) were identical. Briefly, Ss were told that their task was to discover and remember the correct sequence in which to light the 12 lights in the row; that the settings would be called out by number, 1 through 12; that they would have only one chance on each setting; that they were simply to guess for the first trial since the sequence had been determined in an arbitrary, random manner, and systematic sequences of responses were therefore to be avoided; and that they could not expect to do very well at the beginning because of the difficulty of the task. Operation of the apparatus was then demonstrated, with emphasis on the proper use of the stick and rudder bar. Right and wrong signals were demonstrated. The S was required to tell *E*, in his own words, the essential parts of the procedure. Any aspects that seemed

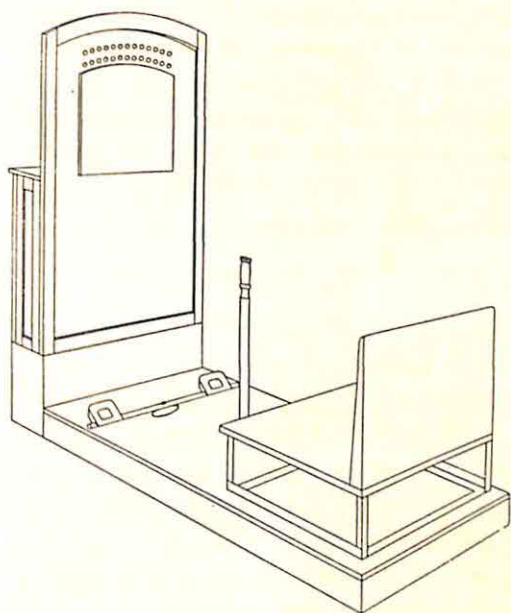


FIG. 1. Drawing of the apparatus.

to be unclear were then reviewed by *E*.

After *E* was satisfied that *S* understood the general procedure to be followed he began the experiment. Light settings were called out as follows: "Setting No. 1," "Setting No. 2," etc. Ss were allowed to proceed at their own pace, although if an *S* was markedly slow he was cautioned to speed up his responses. The trials of 12 responses each averaged approximately 90 sec. A fixed intertrial interval of 30 sec. was used.

E was instructed to watch for any tendency towards systematic response sequences which *S* might show in spite of the strong caution in the instructions. Preliminary experimentation indicated that the most common examples of such sequences were the repetition of a single light position, with the apparent intention of eliminating it, and the progressive movement by one or more light units in one direction along the row of lights. Ss showing such obvious systematic sequences were cautioned by *E* when he was reasonably sure of the sequence, ordinarily after the third response in the series. This caution coming after the emphasis on this point in the instructions was so successful that only seven Ss had to be discarded for having persisted in making such systematic sequences of responses. These were discarded upon examination of the records prior to any scoring or analysis.

The special instructions given to the EU Ss at the end of the first trial, and the entirely different set of instructions given to the CRB group will be found along with the standard instructions (see footnote 3).

³ The detailed instructions used for all conditions have been deposited, along with certain tabulations of data, with the American Documentation Institute. Order Document No. 5281 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.75 for 35-mm. microfilm copies or \$2.50 for 6 × 8-in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Before they left the testing room, Ss were requested to refrain from discussing the experiment. As an additional precaution against the giving of information, Ss who had been tested were kept in a separate waiting room, and in a different building when possible, from those who were yet to be tested. Since Ss at each testing session were drawn from different flights, between which contact was minimal, this possible source of passing information was disregarded.

2. *Experimental Group E*

One subgroup of 150 Ss was rewarded on the second and the eleventh responses throughout the five trials (E-2); the other subgroup was rewarded on the sixth response (E-6). Reward in these stimulus positions was given regardless of the response made. This procedure permitted measurement of error strength for the four stimulus positions *before* response 11 and the four stimulus positions *after* response 2; and for the four stimulus positions before and after response 6. The two subgroups were used to detect possible differences associated with serial position factors.

The primary concern was with the first two trials. However, additional prearranged rewards were given on the third, fourth, and fifth trials. This was chiefly to prevent suspicion and to keep up morale in the subject population in case there were any discussions of the procedure. Little use was made of these additional data.

3. *Experimental Uncertainty Group EU*

This was an added experimental group designed to determine the effect of the so-called "uncertainty" factor (cf. 8). These Ss were treated exactly like the standard experimental (E) Ss throughout the instructions and the first trial. At the end of the first trial, however, they were told that they would be informed during the second trial whenever the light setting they were about to make

was one of those that they had responded to correctly on the first trial. They were then so informed, as the appropriate stimulus was given, but were not told the setting which had been correct on the first trial. It was believed that thus advising the S as to where in the response series he had been right on the first trial would effectively remove any uncertainty he might have on this point and thus eliminate deliberate repetition of responses near the rewarded-response position.

The EU Ss were divided into two subgroups of 150 each according to the position of the rewarded responses. EU-2 Ss were rewarded in the second and eleventh response position; EU-6 Ss were rewarded in the sixth response position. Unlike the E Ss, they were given only two trials.

4. *Control Group C*

This group consisted of 300 Ss who were given no rewards on the first two trials. Records from these two trials constituted the basic control data for comparison with the experimental results. On the third and fourth trials they were arbitrarily rewarded on the second and eleventh responses, making possible the accumulation of additional experimental data. One or two further arbitrary rewards were given on the fifth trial.

The additional 74 Ss from a control group originally designed for a separate purpose were treated exactly like the ordinary controls throughout trial 2 and therefore could be pooled with them for the important analyses based on trials 1 and 2. They differed from the standard controls only in the inclusion of an additional two nonrewarded trials, making a total of four such trials, before the first arbitrary rewards were given in the second and eleventh stimulus positions.

5. *Response-Bias Control Group CRB*

This group was designed to control for possible response biases due to the position of the rewarded response. These Ss helped to test the relevance of Smith's (15) hypothesis that gradients may be produced, as artifacts followed the repetition of a correct response, by the restriction in response range that occurs because of the tendency for Ss to avoid immediate repetition of a response. They also controlled for any pre-experimental biases in the response series, of the kind emphasized by Jenkins and Sheffield (4). Such biases needed to be considered even though the task was designed for relative freedom from them.

CRB Ss were individually matched with the standard E Ss on the basis of the particular light settings made by the latter in the critical, rewarded positions during the first two trials. They were forced to respond in this way by means of adjustable pointers, which were set immediately before each trial. In all other response positions they were allowed to respond as they pleased except for the usual caution against the use of systematic sequences. In order to make this procedure reasonable to the Ss they were told that operating checks were being run on the apparatus. Right and wrong were therefore not involved. Two subgroups of 150 Ss each (CRB-2 and CRB-6) were used, corresponding to the E conditions. They were given only 2 trials.

III. RESULTS

In reading the results, it will be necessary to understand the meaning of several terms that are used repeatedly. "Stimulus position" and "setting number" will be used interchangeably to refer to *serial position* within a trial. That is, "stimulus position 6" refers to that position within

a trial at which E called out "Setting number 6" and S subsequently made his sixth response in the series of 12 responses which constituted a "trial." "Light" or "light setting" refers to the *spatial location* of the particular response S made for any of the stimulus positions within a trial. S could make a light setting from 1 through 12, since there were 12 lights on the panel.

In the experimental groups "error position" generally refers to the stimulus position relative to the reinforced stimulus position. Thus the "+1" error position is the stimulus position immediately after the reinforced stimulus position; the "-1" is the position immediately before, etc. Responses made in the critical (rewarded) stimulus positions (second and eleventh, or sixth) are referred to as "key responses."

In the control groups, for purposes of comparison, error strength was measured before and after certain nonrewarded responses. These responses are therefore also called "key responses," regardless of their stimulus position. Other errors, before and after such key responses, are referred to as "+1," etc., as in the case of the experimental groups.

As earlier described, response strength was measured by the variation, in linear light units, between light settings made on the first two trials at a particular stimulus position. For example, if S responded with the sixth light on trial 1 in the fourth stimulus position, and with the third light on trial 2 in the fourth stimulus position, he was assigned an error difference score of 3 (6 minus 3).

A. *Trial 1-Trial 2 Analysis*

1. *All Ss*

This section presents the data from the 300 Ss of each major group. Marked after-gradients, but no fore-gradients,

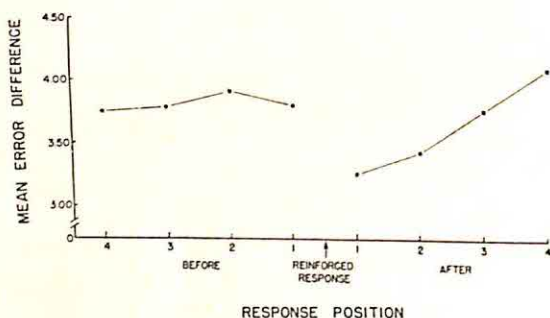


FIG. 2. Error-difference gradients in Group E.

were found for both experimental groups. These are shown in Fig. 2 and 3. Application of a trend analysis, using the technique of analysis of variance (5), showed that both of the experimental after-gradients were significant at well beyond the .01 level of confidence. The F s were 5.46 and 7.22 for E and EU groups respectively; 3.80 is required for 3 and 897 df at the .01 level.⁴

The reversed fore-gradient for Group EU in Fig. 3, which was in a direction opposite to the prediction, was not significant ($F = 2.47$; 2.61 required at the .05 level for 3 and 897 df). F was not calculated for the other curve, since there was no indication of a fore-gradient in the raw score means.

The corresponding curves for the two control groups are shown in Fig. 4 and 5. No evidence of any gradient was found in Group C. The gradient in Group CRB was very nearly statistically significant ($F = 2.58$; 2.61 required at the .05 level for 3 and 897 df). However, this is not as critical an analysis for the CRB Group as the analysis of the records of Ss who repeated the key response (the response made in the same stimulus posi-

⁴ Wherever necessary, Hartley's (2) suggestion for testing homogeneity of variance was applied. In no case was the ratio large enough to cause concern. Norton (12; see also 5) has pointed out that deviations from homogeneity cause smaller losses in accuracy than has been commonly supposed.

tion as the rewarded response for the matched E-group Ss), thus providing an "anchor" for their response biases.

The after-gradients were statistically significant in three of the four experimental subgroups but in none of the control subgroups. These data are available from the ADI (see footnote 3).

2. Repetition of Key Responses

The data were next analyzed according to whether or not the key (rewarded) response was repeated on the second trial. This analysis showed that the after-gradient in the experimental groups was dependent upon the repetition or near-repetition of the rewarded response. This is clearly shown by the three curves in Fig. 6. Here the 300 Ss of Group E have been divided into three subgroups on the basis of the amount of variation in the rewarded response from trial 1 to trial 2. It is apparent from the figure that gradients occurred following both exactly repeated and minimally varied rewarded responses. These were statistically significant at beyond the .01 level of confidence (F 's of 4.93 and 4.01; 3.83 and 3.94 required at the .01 level for 3 and 543, 3 and 135 df , respectively). No gradient occurred following a more extensive variation in the rewarded response. Again, there were no fore-gradients.

Similar results were found for a break-

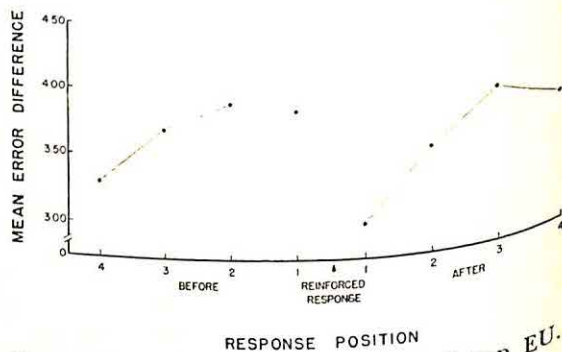


FIG. 3. Error-difference gradients in Group EU.

down of the EU data. With the exception of a slight inversion from position 3 to position 4, a regular after-gradient occurred following repeated rewarded responses. This was significant at well beyond the .01 level ($F = 6.93$; 3.83 required for 3 and 657 df). A perfectly regular gradient was also found following variations of one light unit in the rewarded response, although this was not statistically significant ($F = 1.27$; 2.65 required for .05 level for 3 and 177 df). No gradient was found following greater variations (2 to 11 units) in the rewarded response.

Detailed data for these breakdowns and corresponding data for the forepositions are available from the ADI (see footnote 3). The predicted gradients following repeated rewarded responses occurred for all of the four experimental subgroups, although they were statistically significant only for subgroups E-2 and EU-2. The fact that a highly significant gradient ($P = .01$) was found for Group E-6 following rewarded responses that were varied by one unit suggests that the gradients were not peculiar to an early serial position.

The curves of response strength following repetition of the key response in the two control groups are shown in Fig. 7. The 52 Ss represented are those from the C Group of 300 Ss who repeated either their second or their sixth re-

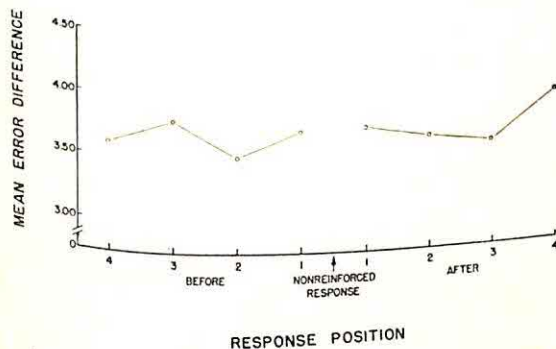


FIG. 4. Error-difference gradients in Group C.

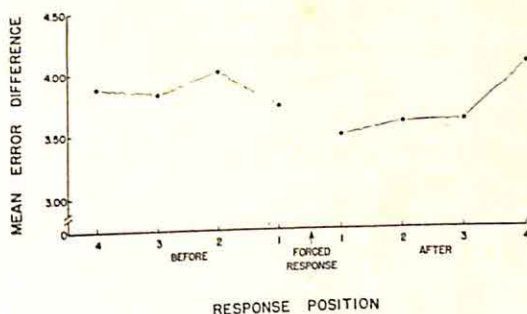


FIG. 5. Error-difference gradients in Group CRB.

sponse. The 182 CRB Ss are those who were matched with E Ss who repeated either their second or their sixth response.

No control gradients occurred. An analysis of variance for the CRB curve, which gave a slight indication of a gradient, showed that it was not significantly different from a horizontal straight line ($F = 1.93$; 2.62 required at .05 level for 3 and 543 df).

Although the gradient for the 182 E Ss who repeated the rewarded response was highly significant (beyond .01 level) and there was no gradient for the 52 C repeaters at the same positions, the direct comparison of the slopes failed to yield a significant difference ($F < 1.00$). The same was true for the comparison of the 182 E Ss with the 182 CRB Ss who repeated at the same positions ($F = 1.74$;

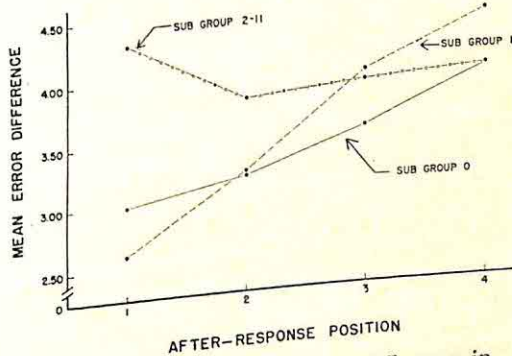


FIG. 6. After-gradients of error difference in Group E as a function of degree of repetition of rewarded response.

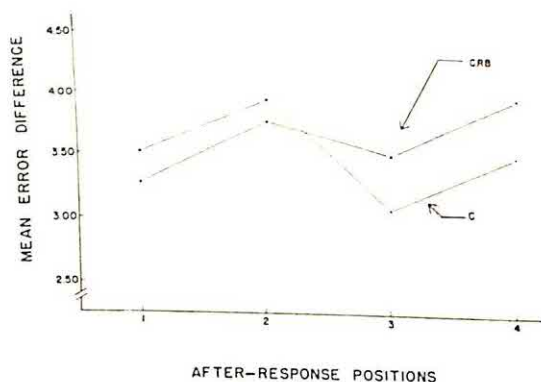


Fig. 7. After-gradients of error difference for control groups following repeated nonrewarded responses.

2.61 required at the .05 level for 3 and 1,086 *df*).

The CRB Ss who exactly repeated their key responses are those to whom the probability-bias hypothesis applies. The gradient for these Ss failed to approach significance. Both of the control curves presented in Fig. 7 were based upon repetitions of responses in positions 2 and 6 only. They thus provide an exact match with the position of the rewarded response in the two experimental groups.

A control curve based upon a larger sample was constructed for gradients following the repetitions of responses in any of the first eight stimulus positions. The use of such repetitions seems to be justified by the finding, noted above, that the experimental gradients occurred in both the second and the sixth positions. The generality of the experimental results over the early and middle serial positions was thus strongly suggested.

A total of 201 Ss who repeated at least one response within response positions 1 through 8 was obtained from the 374 C Ss. Repetitions of responses in the last four positions were not included since they did not permit measurement of four errors following the key response. Since for purposes of statistical analysis each S could be counted only once, it was de-

cided that the first repetition of each S would be analyzed. Results for all of the repetitions were also computed, however, and are available from the ADI (see footnote 3). The response-variation values for this group of 201 C Ss showed some tendency toward an after-gradient because of the low error-difference score for the first error following the repeated response. However, the apparent trend did not approach statistical significance ($F = 1.16$; 2.61 required for the .05 level for 3 and 600 *df*). The inclusion of all of the repetitions of the 201 C Ss who repeated responses within the first eight response positions did not materially change the results.

As additional checks, in order that the first error score be tested independently of the trend test for a gradient, separate *t* tests were run between the mean variation score for the first error and that for each of the other three errors. None of the *t*'s were significant (values of 1.26, 1.75, and .59 for the comparisons with second, third, and fourth error means, respectively; 1.97 required for the .05 level).

For purposes of comparison similar *t*'s were run for the two experimental groups after repeaters. The difference in mean variation scores between first and fourth errors following repeated reward responses was found in each case to be highly significant (*t* for E = 3.52; *t* for EU = 3.76; 2.62 required for .01 level).

A direct comparison of the slopes of the experimental and the above control gradient again failed to show any significant difference ($F = 1.29$; 2.61 required at .05 level for 3 and 1,143 *df*).

B. Analyses Beyond Trial 2

These data were mainly obtained from the use of reward in positions 2 and 11, or 6, on trials 3, 4, and 5 for the 300 C Ss.

They were thus necessarily somewhat contaminated by interference from response biases carried over from trials 1 and 2. The gross results are available from the ADI (see footnote 3).

There was no gradient between trials 2 and 3. Thus reward only on the second trial of a pair did not result in a gradient. There was a very good gradient of error difference for trials 3 and 4. Continuing regular gradients for trials 4 and 5, and for repeaters on trials 2 and 3 in the E group, indicate that the gradients continued past the second rewarded trial. Although no statistical tests were used, owing to the relative impurity of the data, the consistency with which the gradients appeared is of some interest.

C. Tests of Alternative Explanations

A large number of additional analyses were run to test various alternative explanations of the experimental after-gradients.

Thorndike's hypothesis concerned only the presumed strengthening of an S-R connection by contiguity with reward. He assumed that this strengthening, established on the first trial, caused Ss to be more likely to repeat the same error to the same stimulus when the stimulus was presented again on the second trial. It might be argued that the simple demonstration of differences between experimental and control gradients is sufficient to validate Thorndike's hypothesis, since the control gradients eliminate the usual guessing-sequence and probability-bias hypotheses.

However, it is possible that reward acts in some fashion other than the "spread" or "scatter" postulated by Thorndike (17). Reward might create a stronger intratrial relationship between responses on the first trial, and also stronger intratrial relationships on the second trial.

This would result in a stronger apparent intertrial relationship between first errors, yet would not be the kind of *direct* S-R strengthening with which Thorndike was concerned. For example, if the key response were the seventh light, and reward increased the probability that the *following* response would be the eighth light, then the probability that the error difference score would be zero (S responding with eighth light at the first after-error position on both trials) would be increased, independently of any actual strengthening of the S-R connection between environmental stimuli and the choice of the eighth light.

The analyses reported here were run to see whether an increase in intratrial relationships between responses actually occurred as a side-effect of reward; and, if so, whether this kind of response bias was sufficient to account for all of the observed intertrial strengthening. Only gradients following repetition of the key response are considered, since only in that case were significant gradients found. These analyses should therefore give some information about the degree of confidence that could be accorded the conclusion that the after-gradients were not entirely artifactual.

1. The Alternation Hypothesis

If Ss tended to respond first to one side, then to another, in this apparatus, gradients could be created artifactually by this tendency. The key response if repeated would serve as an anchor point on the same side on both trials. Alternation of the next response to the opposite side on each trial would then result in a lowered intertrial difference score for that response. This follows from the fact that if the response in any serial position was made to the same side on both trials, the maximum possible error difference was 5; if the response was made to different sides on the two trials, the maximum possible difference was 11. However, the alternation tendency could not explain why the experimental gradients following repetitions were stronger than the control, unless the alternation tendency was stronger in the experimental group.

TABLE 1
PERCENTAGE OF ALTERNATION AND MEAN
ERROR-DIFFERENCE SCORES AFTER
REPEATED KEY RESPONSES

Group	N	Trial 1	Trial 2	Error Differences
E-2	85	77.6**	67.1	2.55
E-6	97	57.7	55.7	3.37
EU-2	92	68.5	68.5	3.01
EU-6	128	78.9**	65.6	2.82
C	201	59.7	57.2	3.24
CRB-2	85	80.0**	63.5	3.41
CRB-6	97	80.4**	69.1*	3.05

* Significant at .05 level, compared with Control value.

** Significant at .01 level.

Table 1 shows this analysis for the first errors. The first chance repetition at any position from 1 through 8 was used for the C group (comparable data for all Ss are available from the ADI; see footnote 3).

Two of the four differences between experimental and control groups were significant at the .01 level on trial 1, and both of the differences between the CRB and C groups were significant at the same level. On trial 2, however, none of the experimental percentages were significantly higher than the corresponding control percentages.

The low alternation percentages for E-6 Ss corresponded to weak error reinforcements, but the highest alternation percentages, in the CRB group, also corresponded to weak error reinforcements. Thus alternation alone was apparently not sufficient to produce gradients in these groups. The last column of Table 1 shows these mean first-error difference scores.

Since these analyses were performed exclusively on Ss who repeated the key response, concurrence of either alternation (A) or nonalternation (N) on the two trials resulted in first after-errors on the same side of the central-rest position. Reversal of alternation or nonalternation condition on the second trial resulted in first after-errors on opposite sides of central rest. Accordingly, the Ss were divided into four classes on the basis of their alternation or nonalternation on trials 1 and 2. These were AA, NN, AN, and NA. The first letter designates whether or not S alternated sides from key response to first error on trial 1, the second letter on trial 2. For example, a S who made the key response on the left side on trial one, then the first error to the right on the same trial, would be classified with first letter A. As a repeater, he necessarily repeated the key response to the left on the second trial. Now, if

he responded to the right on the trial 2 first error, he was classified in AA.

The gradients should not have differed significantly within any of the alternation classes if alternation gave rise to the gradients. That is, once the groups have been equalized with regard to alternation, the experimental treatments should not give rise to any significant differences.

Table 2 shows this breakdown. Classes AA and NN were combined on the basis of the fact that the response tendency was consistent on the two trials, resulting in first-error responses to the same side on both trials. In both classes, the error-difference score would tend to be decreased. A *t* test was run between the difference scores for C and E groups, and the difference was significant at the .02 level. However, when EU classes AA and NN were combined, the mean difference score for EU was exactly the same as that for C.

Classes AN and NA were similarly pooled. The *t* for C versus EU was significant at the .03 level, while that for C versus E was not significant.

Thus, when the alternation tendency was controlled by classifying the responses as outlined above, two of the four similar experimental groups showed significantly smaller difference scores after reward. Considering only these four comparisons, the chances of getting two of four *P*-values as small as even .05 by chance is only .015. This calculation gives some indication that there was still a difference between groups after the alternation tendency was controlled.

The same type of analysis into groups by response tendencies on trials 1 and 2 was performed for the CRB group, as shown in Table 2. No statistical tests were made of the differences between C and CRB groups, since inspection of the table clearly indicated that the differences would not be significant.

The alternation analysis gives some interesting information about subgroup E-6. This group showed a very weak error-reinforcement effect with its relatively high difference score for the first after-error of 3.37. This value was greater

TABLE 2
MEAN FIRST-ERROR DIFFERENCE SCORES AFTER
REPEATED KEY RESPONSES FOR
ALTERNATION CLASSES

Group	Same sides,			Different sides,		
	N	AA & NN	<i>t</i> ^a	N	AN & NA	<i>t</i> ^a
E	104	1.29	2.51*	77	5.31	1.19
EU	141	1.69	.00	81	4.98	2.23*
C (I-II)	148	1.69		92	5.74	
CRB	115	1.53		67	6.12	

^a E and EU tested against C.

* Significant at .03 level.

TABLE 3

PERCENTAGE OF RESPONSES TO SAME AND DIFFERENT SIDES AFTER REPEATED KEY RESPONSES

Item	Group				
	E	EU	C (1-8)	C (1-11)	CRB
Total N	182	220	201	242	182
% Same Sides (AA & NN)	57.7	63.2	60.7	62.0	63.2
% Different Sides (AN & NA)	42.3	36.8	39.3	38.0	36.8

than that for any of the other experimental groups, and also greater than the control value of 3.22. However, when the group was broken down into components AA, AN, etc., none of the E-6 means were greater than those of the corresponding control groups. In other words, the alternation tendency worked against the tendency for errors to be reinforced in this group. The fact that the over-all score for E-6 was higher than for C is thus explained by the fact that the percentage of the group which either alternated or did not alternate consistently on the first two trials (AA or NN) was significantly smaller than in the control group ($\chi^2 = 6.89$, $P < .02$ for 1 df).

This brings up the possibility that the over-all gradients following repetitions in the experimental groups might be explained by a larger number of experimental Ss responding to the same side with the first error on trials 1 and 2. This would artificially reduce the difference score. Such a tendency would be shown by a larger number of experimental than control Ss appearing in classes AA and NN. Table 3 shows the percentages of Ss in each group who responded to the same and to different sides on the two trials.

The over-all difference among the proportions of E, EU, and C groups in classes AA and NN was not significant ($\chi^2 < 1.00$ for 1 df). This verifies what one might conclude from a gross inspection of the table: there does not seem to be any significant difference in the tendency to respond to the same side on both trials with the first error response. The over-all differences among groups are not explicable in terms of differential numbers of Ss who responded to the same side on both trials.

2. The Central-Rest Tendency

Even though it did not appear likely that there was any group difference in the tendency to respond to the same side with the first errors on

the two trials, there remained a possibility that the reward affected the distribution of responses. A tendency was noted for Ss to restrict their response range following reward; this restriction would reduce the range of difference scores and might therefore have produced an artificial gradient.

In order to test for this possible central-rest tendency, responses were arbitrarily divided into "central" locations (light settings 4 through 9) and "peripheral" locations (light settings 1 through 3 and 10 through 12).

Table 4 shows the percentages of first errors following repetitions that were in central locations. Three out of four obtained experimental values were smaller than the control values on trial 1; EU-2 was significantly smaller. In spite of the fact that E-6 Ss had a significantly larger percentage in central locations, they did not show a significant error-difference gradient.

On trial 2, only EU-2 had a significantly larger value than the control. This shift from significantly smaller to significantly larger in group EU-2 was paralleled to some extent by a corresponding shift in EU-6, although the differences were not significant for this group. Altogether, it does not seem that the difference in percentages in central locations could have produced the gradients.

Further analyses were then run in order to compare the difference-score gradients following repetitions of central key responses with the gradients following repetitions of peripheral key responses. Central responses were again 1 to 3 units from the central-rest position, peripheral responses 4 to 6 units away. Table 5 shows these gradients and their *P*-values.

There were significant gradients following the repetitions of a central key response in the ex-

TABLE 4
PERCENTAGE OF CENTRAL FIRST-ERROR RESPONSES AFTER REPEATED KEY RESPONSES

Group	N	Trial 1	Trial 2
E-2	85	55.3	63.5 ^c
E-6	97	68.0 ^{a,d}	64.9 ^d
EU-2	92	44.6 ^a	76.1 ^{b,d}
EU-6	128	50.0	67.2 ^d
C	201	56.7	59.7
CRB-2	85	41.2 ^b	52.9
CRB-6	97	52.6	51.5

^a Significantly different from C value at .05 level.

^b Significantly different from C value at .01 level.

^c Significantly different from 50 per cent at .05 level.

^d Significantly different from 50 per cent at .01 level.

Bureau Ednl.
Filing No.
Dated
No.

TABLE 5
MEAN ERROR-DIFFERENCE SCORES AFTER CENTRAL AND PERIPHERAL
REPEATED KEY RESPONSES

Group	N	After-position				p
		1	2	3	4	
E						
Central	108	2.75	3.08	3.71	4.30	.01
Peripheral	74	3.32	3.43	3.32	3.45	—
EU						
Central	135	2.69	3.70	4.10	3.57	.01
Peripheral	82	3.26	3.23	3.68	4.04	—
C (0 & 1) ^a						
Central	60	4.12	3.78	3.68	4.38	—
Peripheral	22	3.09	2.45	4.09	3.27	—
CRB						
Central	108	3.44	3.59	3.55	3.63	—
Peripheral	74	2.95	3.82	3.31	4.18	—

^a In order to increase the small N for control repetition in positions 2 and 6, Ss with a difference score of one were included.

perimental groups, but no other significant gradients. There was an interesting tendency for first errors after peripheral key responses to be stronger than after central key responses in the control group. This is the reverse of the situation in the experimental groups. This tendency was checked by tabulating the first-error strength after central and peripheral key responses over all stimulus positions for the control group. Seven of the eight positions checked showed a stronger peripheral error, with position five alone showing a stronger central error (data are available from ADI; see footnote 3). Thus this effect in the control group does not appear to have depended on the serial position in which the key response occurred.

Since there were these differences between the gradients following central and peripheral key responses, it is not true that a simple restriction in response range following any reward is adequate to explain the gradients. The remaining possibility is that there was a gradient of restriction of response range following reward of a central response, but not of a peripheral response.

This possibility was also checked. A different kind of gradient was calculated. This was a gradient of distances of the successive error responses from the central-rest position. The absolute distance from central rest, irrespective of direction, was used for this analysis. The error-difference score would tend to be reduced if responses were near the resting position, irrespective of whether the responses were to the left or right.

Table 6 shows that central-rest gradients were found in the two central subgroups, which had significant error-difference gradients. There was also a significant gradient in the EU group following peripheral key responses, but the gradient was in a reversed direction; that is, the response range was *less* restricted immediately following the key response than it was farther from it. These results suggest that there was some tendency to make the first after-error at about the same distance from the center as the rewarded response.

In general, the results were at least consistent with a hypothesis that the gradients were produced by a gradient in distance of the response from the central-rest position. It is impossible to evaluate from this analysis alone whether or not the central-rest tendency did actually produce the gradients. In the preceding analysis, the concern has been with the distribution of responses following reward. There was still a question of whether or not there was any tendency for errors to be strengthened directly, independently of such distributional artifacts. A method was needed which would be relatively independent of the distribution of responses following reward, yet which would reveal strengthening of errors. This method is next described.

3. Intertrial Correlational Analyses

The product-moment correlation was used to measure the degree of association between first- and second-trial first after-errors. This measure was chosen because the distribution of *r* is fairly

TABLE 6
MEAN TRIAL 1 CENTRAL-REST SCORES AFTER CENTRAL AND PERIPHERAL
REPEATED KEY RESPONSES

Group	N	After-position				P
		1	2	3	4	
E						
Central	108	2.05	3.33	3.30	3.89	.01
Peripheral	75	3.52	3.23	3.20	3.12	—
EU						
Central	135	3.18	3.43	3.94	3.59	.01
Peripheral	82	3.86	3.15	3.12	3.24	.05
C(0 & 1) ^a						
Central	60	3.48	3.40	3.50	3.27	—
Peripheral	22	3.59	3.54	3.41	3.23	—
CRB						
Central	108	3.78	3.93	3.52	3.77	—
Peripheral	74	3.68	3.63	3.19	3.30	—

^aIn order to increase the otherwise small N, Ss with an error-difference score of one were included.

insensitive to deviations from the normal bivariate distribution in the parent population (1, 13). The observed r , however, might have been produced by the relationship between the key response and first error on trial 1, in conjunction with a similar relationship on trial 2. These relationships can be partialled out of the final r through the use of partial correlation.

Correlations were computed between trial 1 and trial 2 first after-errors after key response,

and between key response and first after-error on trial 1 and 2. Again, the analysis was only for Ss who repeated the key response. Partial correlations were derived from these three correlations, so that the intratrial correlations were partialled out of the final correlation, which became a pure measure of intertrial correlation. Table 7 shows the results.

All the intratrial correlations were negative, although several were not statistically significant.

TABLE 7
INTER- AND INTRATRIAL CORRELATIONS AND PARTIAL CORRELATIONS FOR
FIRST ERRORS AFTER REPEATED KEY RESPONSES

Group	N	Intertrial r	Trial 1 Intratrial r	Trial 2 Intratrial r	Partial Intertrial r
E					
After 2	85	.31 ^a	-.46 ^b	-.18	.26 ^a
After 6	97	-.08	-.22 ^a	-.17	-.12
Total	182	.13	-.33 ^b	-.18 ^a	.07
EU					
After 2	92	.30 ^a	-.38 ^b	.23 ^a	.21 ^a
After 6	128	.28 ^a	-.52 ^b	-.25 ^b	.17 ^a
Total	220	.28 ^b	-.46 ^b	-.21 ^b	.21 ^b
C	201	.11	-.01	-.01	.11
CRB					
After 2	85	.11	-.53 ^b	-.24 ^a	-.02
After 6	97	.36 ^b	-.47 ^b	-.25 ^c	.28 ^b
Total	182	.29 ^a	-.50 ^b	-.25 ^b	.20 ^b

^aSignificantly different from zero at the .05 level.

^bSignificantly different from zero at the .01 level.

TABLE 8
INTERTRIAL CORRELATIONS FOR FIRST ERRORS
AFTER CENTRAL AND PERIPHERAL
REPEATED KEY RESPONSES

Group	Central		Peripheral	
	N	r	N	r
E				
After 2	55	.39 ^a	30	-.30
After 6	53	-.03	44	-.07
Total	108	.20 ^b	74	.07
EU				
After 2	59	.36 ^b	33	.21
After 6	79	.21	49	.31 ^a
Total	138	.28 ^b	82	.25 ^a
C (1-8)				
Total	143	.02	58	.28 ^a
CRB				
After 2	55	.34 ^b	30	.02
After 6	53	.23	44	.40 ^b
Total	108	.28 ^b	74	.24 ^a

^a Significantly different from zero at .05 level.

^b Significantly different from zero at .01 level.

The negative correlations reflect the alternation tendency. In the control group, the intratrial correlation reflecting this tendency was negligible.

The difference between the control intratrial r for trial 1 and the r for the other groups was significant at the .05 level, or better, for groups E-2, EU-2, EU-6, CRB-2, and CRB-6. On trial 2, the same significance was found between control and EU-2, EU-6, CRB-2, and CRB-6. Since only Ss who repeated were considered in this analysis, the differences were not produced by any artifacts due to differential repetition. Reward, as in the E and EU groups, or forcing, as in the CRB group, thus seems to increase the operation of response biases of the kind shown in intratrial correlation.

After the intratrial effects had been ruled out, three out of the four experimental subgroups still showed significant partial r s, indicating "pure" intertrial associations between first errors. However, a negative r was obtained for E-6, corresponding to its weak error-reinforcement. Thus, when this group was pooled with E-2 to get the total E, the correlation was actually smaller than that for the C group, though not significantly so. It is difficult to see why there should be such a marked difference between the second and sixth position subgroups in this E group. It was pointed out earlier that the alternation tendency worked against any error reinforcement in this group, but this tendency needs explaining also. In the absence of any evident

explanation for this effect, an interpretation in terms of sampling variation cannot be rejected.

The control intertrial r and its partial r were both positive, but were not significantly greater than zero. The differences between the control partial r and the other partial r s were tested for significance, using the z' transformation method (6). None of the differences was significant. Thus the partial correlation technique failed to reveal any pure intertrial effects that differed significantly from E to C groups. It is still true that the obtained values of r in the experimental groups were significantly larger than zero, with the exception of the E-6 group.

An interesting incidental finding is that the intratrial r s were without exception smaller on trial 2. It may be that nonreward of the first after-error tended to weaken the tendency to make that response to the stimuli arising from the reward of the preceding key response.

Similar correlations were run between second, third, and fourth errors following the key response. The correlations were generally nonsignificant.

A further attempt was made to evaluate the intertrial correlation between first after-errors in terms of the central or peripheral position of the key response. Table 8 shows the results.

In general, the results were what might be expected on the basis of the central-rest analysis. The control groups showed a higher correlation following key responses in peripheral positions, while the experimental groups showed more significant results following key responses in central positions.

4. "Mirroring" Analysis

An apparent "mirroring" effect was observed in the experimental Ss (E and EU combined). This was the tendency for Ss whose key response and first after-error were on opposite sides of the center of the apparatus to respond about the same distance from the central-rest position on the two successive responses. For example, a S who responded with a no. 10 light for the rewarded response has gone 4 units to the right of the resting position. If he alternated sides on the succeeding response, he would tend to respond approximately the same distance to the left, that is close to the no. 3 light (4 units to the left of the central-rest position).

Evidence for this tendency is given by the correlations in Table 9. These correlations are in terms of the absolute distance from the central-rest position. There was a significant relationship between the key response and first after-error in the experimental groups on trial 1. In no case was a control correlation significant. These mirroring effects largely disappeared on the second trial.

This analysis is suggestive with regard to the failure to find significant experimental gradients following peripheral responses. As a result, presumably, of the mirroring tendency, a significantly larger proportion of such experimental Ss responded to the periphery on the opposite side following a rewarded key response on trial 1. Out of 31 Ss in Group C that had peripheral key responses and alternated their succeeding response, only 10 responded to the peripheral area. In the combined E and EU groups, 67 of 108 such responses were to the opposite periphery. This difference was highly significant ($\chi^2 = 8.72$, $P = .01$ for 1 *df*). The total number of peripheral first-error responses that followed peripheral key responses was also significantly higher for the experimental than for the control Ss (85 of 156, compared with 22 of 58; $\chi^2 = 4.64$, $P = .05$, for 1 *df*).

These differences are important because when the trial 1 first-error response was to the periphery, the intertrial difference score expected would be high unless there was again a mirroring effect on trial 2. Since the mirroring effect dissipated on trial 2, with a slightly higher proportion of control than experimental peripheral first errors, one would expect higher difference scores for experimental Ss. Thus the response biasing in terms of mirroring worked against intertrial error-difference gradients in the experimental groups. This point is of interest in indicating that response biases do not necessarily favor the production of gradients, or strong first errors, following repeated rewarded responses.

The strength of the mirroring effect, as indicated by the first-trial correlations in Table 9, suggests the possibility that it might be an important contributor to what has been described as the central-rest tendency in the experimental Ss. The greatest contribution to this tendency was made by Ss whose central first errors were alternated following central key responses. Mirroring and central-rest factors are obviously confounded in such central-central alternations.

In order to determine whether a genuine central-rest tendency remained when mirroring as a factor was eliminated, a check was made involving only nonalternating first errors. On the assumption of an independent central-rest tendency following reward, it was predicted that the proportion of nonalternating Ss shifting from peripheral to central would be greater than the proportion shifting from central to peripheral. This prediction was confirmed ($\chi^2 = 2.74$, $P = .10$ for 1 *df*, on trial 1; $\chi^2 = 8.16$, $P = .01$ for 1 *df*, on trial 2). The comparable checks for Group C and CRB failed to show the same effect, three of the four proportions being in the opposite direction and the fourth being very slight and nonsignificant ($\chi^2 < 1.00$).

It is thus concluded that the present study

TABLE 9
CORRELATIONS BETWEEN CENTRAL-REST SCORES
OF KEY RESPONSES AND ALTERNATING
FIRST ERRORS ("MIRRORING" EFFECT)

Group	First Trial					
	AA		AN		Combined	
	N	r	N	r	N	r
E	78	.15	44	.45**	122	.26*
EU	115	.46**	49	.32*	164	.41**
C	93	.09	47	-.09	141	.00
CRB	100	.11	49	.12	149	.10

Group	Second Trial					
	AA		NA		Combined	
	N	r	N	r	N	r
E	78	.21	33	.16	111	.19*
EU	115	.03	32	.25	147	.08
C	93	.15	44	-.24	137	-.02
CRB	100	-.01	23	-.05	123	.10

* Significant at the .05 level.

** Significant at the .01 level.

does provide independent evidence for the central-rest tendency as a side-effect of reward. Evidence for mirroring independent of the central-rest factor is of course also provided by the significant peripheral-peripheral shifts described above.

The various intratrial comparisons made in the present section were based on summary data tabulations available from the ADI (see footnote 3).

5. Combined Alternation and Central-Rest Analysis

These analyses were an attempt to further analyze the gradients in terms of combinations of response-biasing effects. The first analysis concerns the percentage of each of the major groups within categories which tend to reduce the error-difference scores and within categories where the bias is toward increasing the error-difference scores. Groups AA and NN are again combined, since the first error was on the same side on both trials, and AN and NA are combined since the first error was to different sides of the central-rest position.

As is evident in Table 10, there was a relatively small difference in the percentage of each major group within each category. This indicates that the observed differences in gradients did not depend upon this difference in percentage of the group whose responses were biased toward low difference scores. For example, the percentage of

TABLE 10

PERCENTAGE OF Ss IN VARIOUS CENTRAL-PERIPHERAL AND ALTERNATION CATEGORIES

Group	N	Central			Peripheral		
		Same Sides (AA, NN)	Different Sides (AN, NA)	Total	Same Sides (AA, NN)	Different Sides (AN, NA)	Total
E	182	36.2	23.1	59.3	21.4	19.3	40.7
EU	220	40.0	22.7	62.7	23.2	14.1	37.3
C	201	42.8	28.3	71.1	18.4	10.5	28.9
CRB	182	13.2	46.2	59.4	7.1	33.5	40.6

the C group in the AA, NN, central category was larger than for either of the other groups, and this category should produce the smallest difference scores on the basis of the central-rest tendency and the alternation tendency per se.

The first-error scores actually found for each of the subgroups are shown in Table 11. There an interesting difference was revealed between the C and E groups. In the C group, the peripheral errors were strongest. In the E groups, the central errors were strongest, following reward. To check the effect further, special control data composed of repetitions by chance, not rewarded, in the E groups, were tabulated. As shown in Table 11, these repetitions were also followed by stronger peripheral errors.

Inspection of the distribution of central and peripheral first errors following central and peripheral repeated key responses suggests an interpretation of this result. On trial 1 the ratio of central to peripheral first errors was greater following a peripheral key response for the control Ss, and was greater following a central key response for the experimental Ss. This difference

in proportions approached significance at the .05 level ($\chi^2 = 3.23$; 3.84 required for 1 df). A high proportion of central first errors would produce a natural bias towards lower intertrial error-difference scores, as has been shown earlier.

The greater tendency for peripheral key responses to be followed by central first errors in the control Ss is explainable on the basis of the relatively lower number of peripheral-peripheral alternating relationships (such as were apparently produced by the mirroring tendency in the experimental Ss), and the consequent opportunity for more or less random first-error responses to occur more frequently therefore in the central areas.

Table 11 also shows the *t* tests for the various combinations of first errors. Following central key responses, two of the four subgroup differences were significant. Both the combined analyses after central key responses approached significance. Since certain of the *t*s approached the accepted level of significance, their *P* values are given even though they were slightly greater than .05. Thus an analysis which controlled for

TABLE 11
MEAN FIRST-ERROR DIFFERENCE SCORES FOLLOWING CENTRAL AND PERIPHERAL KEY RESPONSES, ALSO CATEGORIZED BY ALTERNATION CLASSES

Group	Same Sides (AA & NN)			Different Sides (AN & NA)			Combined		
	Cent.	Periph.	Total	Cent.	Periph.	Total	Cent.	Periph.	Total
E	1.26	1.33	1.29	5.12	5.54	5.31	2.76	3.32	2.99
EU	1.50	2.00	1.69	4.76	5.32	4.98	2.68	3.26	2.90
C	1.73	1.45	1.63	5.80	5.38	5.73	3.39	2.87	3.24
E (C) ^a	1.64	1.58	1.61	5.94	5.10	5.64	3.92	3.18	3.63
EU (C) ^a	1.70	1.69	1.69	5.95	5.54	5.81	3.33	2.77	3.09
Group <i>t</i> tests									
E vs. C	2.12*	.47		1.44	.03		1.83	.82	
EU vs. C	1.08	1.84		2.71**	.53		2.37*	.74	

^a E (C) and EU (C) refer to first errors after chance repetitions in the E and EU groups.

* Significant at .05 level.

** Significant at .01 level.

both alternation and central-rest factors still revealed differences at least approaching significance, and in some cases reaching it.

None of the comparisons showed a clearly significant difference following repeated peripheral responses. However, differences following peripheral responses were biased in the experimental groups as a function of the mirroring effects, as pointed out in the preceding section; they cannot be regarded as constituting a fair test of the spread hypothesis.

IV. DISCUSSION

It is not possible, on the basis of the present results, to make a single simple and conclusive positive statement concerning the reality of error gradients after all of the possible artifacts have been removed. The critical experimental-control comparisons were not significant, even though the experimental gradients by themselves were generally significant for repeaters and the comparable controls were not.

On the other hand, a great deal of positive evidence was obtained in the various control analyses involving the first after-errors alone. For example, two of the four highly refined alternation-analysis results were significant at beyond the .05 level for the experimental subgroups. Three out of the four partial correlations were also significant for the experimental subgroups, in analyses that eliminated the distributional artifacts such as the central-rest tendency. Two of the four direct experimental-control comparisons in the final analysis combining alternation and central-rest factors were significant for the central rewarded responses. Moreover, it was shown that the puzzling failure of the experimental gradients to occur following repetitions of peripheral rewarded responses could well have been an artifact of the bias introduced by the "mirroring" effect. Of the major experimental-control first-error comparisons, only the partial *rs* failed to show some significant differences. Even

here, three of the four subgroup differences were in the predicted direction.

The present experiment provides two additional important findings. One is that reward produces, as a side-effect, stronger intratrial relationships than nonreward. The second is that the experimental gradients seemed to depend on the repetition or near-repetition of the rewarded response.

On the assumption that the control analyses produced sufficient positive evidence to justify continued concern with the problem, an explanation of the Thorndike gradients in terms of both the reward and the repetition of key response factors was attempted. This hypothesis will be called the *serial response-response reinforcement hypothesis*.

According to the serial R-R reinforcement hypothesis, the repeated rewarded response plays a dual role in producing a gradient. It serves, first, to reinforce, through its reward (Thorndikian) action, the responses that follow. This takes account of the presumed superiority of repeated-rewarded over repeated-nonrewarded responses in producing gradients. Second, it serves as a constant element in the stimulus complex. This takes account of the apparent necessity for repetition of the rewarded response. Thus the S-R bonds, if Thorndike's terminology is used, are between successive responses, acting also as stimuli, as well as between environmental stimuli and responses, as Thorndike assumed.

This hypothesis differs from the original Thorndikian "spread" hypothesis and also from Tolman's (18) suggestion of reinforcement of blocks of responses in that it refers only to the *after-gradient* and that it assumes the dependence of this gradient upon the *repetition*, or near-repetition, of the rewarded response. The term "serial" is appended to indicate

specificity of reference to that kind of sequential response situation.

The present experiment affords some evidence which supports this hypothesis. The intratrial correlations were without exception higher following reward than following nonreward. There were stronger alternation tendencies following reward than were found in the control group. Central-rest gradients were found only following reward. All these lines of evidence indicate that following repeated rewarded responses, the responses of the *S* become more predictable. This is not simply a function of the fact that the response was repeated, since the key response was also repeated by the control *S*.

The CRB group is especially interesting in this respect. Here the *S* was forced to make a given response on both trials. The intratrial *rs* were especially high for this group; three of the four were significant at beyond the .01 level, and the fourth at beyond the .05 level. Yet there was no evidence of a gradient. This indicates the necessity of the Thorndikian type of strengthening to produce a gradient. Thus the CRB group gives further evidence that biasing of response sequences is not always sufficient to produce error gradients.

More direct support for the serial R-R hypothesis, which was admittedly *ad hoc* in the present study, has come from two subsequent experiments (7, 10) in which direct experimental-control comparisons have produced significant differences.

Certain incidental results of the present experiment have interesting implications and offer suggestions for further research.

The *central-rest tendency* consisted of a significant tendency for *Ss* to reduce the magnitude of their response (amplitude of stick excursion from central-rest position) immediately following re-

ward of a central response. Theoretically, this might be interpreted as indicative of an increased caution, or perhaps of increased relaxation, as a function of reward.

The *mirroring effect* consisted of a significant tendency for *Ss* to match their rewarded response, in terms of distance of movement from the central-rest position, when the first after-error was made on the alternate side from the rewarded response. This effect was most prominent on the first trial. Theoretically, it offers some interesting support for the automatic effects of reward, since it can be explained in terms of an immediate repetition of a response following reward, if the response is now considered to be a movement of a certain magnitude from the central-rest position.

Practically, these effects may be of interest as possible biasing factors in the operation of complex controls.

Finally, Group CRB is of interest in that these *Ss*, while generally not offering significant gradients, did show certain indications of "reinforcing" influences. Theoretically, the question here raised is the extent to which "forcing" of responses, as used in this control group, may take on "reinforcing" properties. The problem is related to the old theoretical one of the role of "exercise." This kind of forcing technique seems to offer certain advantages for the further investigation of the problem.

V. SUMMARY

This study was designed to test the Thorndikian spread-of-effect hypothesis that errors are differentially strengthened as a direct function of their nearness to rewarded responses. More specifically, it was designed to determine whether statistically significant gradients of response strength would still be found after removal of the guessing sequence and probability-bias factors by control groups directly comparable to the experimental groups. Such direct experimental-control comparisons have not hitherto been made in the investigations purporting to show that gradients can occur in the absence of the usual reward.

Four main groups of 300 *Ss* each were used. *Ss* performed individually on a modified complex coordinator apparatus (11).

There were two experimental groups. In the standard experimental group Ss received one or two prearranged rewards, regardless of the particular responses made in the critical positions. The second experimental group included a control for "uncertainty" on the test trial. The standard control group consisted of 300 Ss given no reward for the first two trials. An additional 74 Ss were added to this group for most of the analyses. The second control group consisted of Ss who were individually matched with the standard experimental Ss. In the positions where experimental Ss were rewarded, these Ss were forced to make the same response. This group was designed to provide an additional check of the response-bias factors. The Ss were told that right and wrong were not involved since operational checks were being made on the apparatus.

The predicted after-error gradients were found in the two groups given reward. No such gradients were found in the control groups. No fore-gradients were found. The experimental gradients were found to depend upon the repetition, or near-repetition, of the rewarded response. Experimental gradients following repetition were significant, and control gradients following nonrewarded but repeated responses did not occur; but there were no significant differences between experimental and control gradients. There were, however, a number of significant differences between the strengths of the first after-error responses.

Possible distributional artifacts were controlled by using a correlation technique, which is relatively insensitive to changes in distribution of errors. These results tended to agree with the results of the analysis of variance of the gradients. The experimental correlations tended to be significant, and the control not sig-

nificant, but the differences between experimental and control correlations were not significant.

Partial correlations were used to take out the intertrial correlations which might occur as a result of a concurrence of intratrial correlation between key responses and first error on the two trials. This was important because the intratrial relationships between the key responses and first after-errors were increased following reward or forcing, in nearly every case, compared to the observed relationship following repeated key responses in the control group. The partial correlations showed the same results as the gross intertrial correlations: three out of four significant partial r s in the experimental subgroups, no significant partial r s in the control group, and no significant experimental-control differences.

Although the apparatus was designed to be free of pre-experimental response biases, some interesting biases appeared in the experiment. First, there was a tendency for Ss to respond to alternate sides of the apparatus. Second, experimental Ss tended to respond closer to the center of the apparatus following reward of a central response. Third, those Ss who responded to alternate sides of the apparatus on the key response and the first after-error tended to respond about the same distance from center on the two successive responses. This "mirroring" effect was most apparent on the first trial. All these response biases were greater following reward.

Refined control analyses were done, controlling for both alternation and central-rest tendencies. Some of these comparisons of first after-errors reached significance, and others approached it. Significant differences were found only following repeated *central* key responses. The mirroring tendency worked *against*

gradients in the experimental groups following repeated peripheral key responses. This illustrates the important fact that response biases do not always work for the production of gradients.

It was concluded that no single simple and conclusive positive statement could be made concerning the reality of the residual experimental gradients after removal of the intratrial response biases. Nevertheless, there was sufficient positive

evidence to justify continued concern with the problem. This evidence suggested a hypothesis, called the *serial response-response reinforcement hypothesis*, which postulates a dual role for the rewarded response. According to this hypothesis, it serves both as a constant element in the stimulus situation, thus increasing intratrial relationships, and as a direct strengthener of errors in the Thorndikian manner.

REFERENCES

- CHESIRE, LEONE, OLDIS, ELENE, & PEARSON, S. Further experiments on the sampling distribution of the correlation coefficient. *J. Amer. statist. Ass.*, 27, 1932, 121-128.
- HARTLEY, H. O. The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, 1950, 37, 308-312.
- JENKINS, W. O., & CUNNINGHAM, LETA M. The guessing-sequence hypothesis, the 'spread of effect,' and number-guessing habits. *J. exp. Psychol.*, 1949, 39, 158-168.
- JENKINS, W. O., & SHEFFIELD, F. D. Rehearsal and guessing habits as sources of the 'spread of effect.' *J. exp. Psychol.*, 1946, 36, 316-330.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.
- MARX, M. H. A confirmation of a modified Thorndikian spread-of-effect hypothesis in a normal multiple-choice learning situation. *Amer. Psychologist*, 1953, 8, 508.
- MARX, M. H. Spread of effect: a critical review. *Genet. Psychol. Monogr.*, 1956, 53, 119-186.
- MARX, M. H., & BUNCH, M. E. New gradients of error reinforcement in multiple-choice human learning. *J. exp. Psychol.*, 1951, 41, 93-104.
- MARX, M. H., & GOODSON, F. E. Further gradients of error reinforcement following repeated reinforced responses. *J. exp. Psychol.*, 1956, 51, 421-428.
- MELTON, A. W. (Ed.) *Apparatus tests*. AAF Aviation Psychol. Program Res. Rep. No. 4, 1947.
- NORTON, D. W. An empirical investigation of some effects of nonnormalities and heterogeneity on the F-distribution. Unpublished Ph.D. thesis, State Univer. of Iowa, 1952.
- PEARSON, E. S. The test of significance for the correlation coefficient, *J. Amer. statist. Ass.*, 1931, 26, 128-134.
- SHEFFIELD, F. D. "Spread of effect" without reward or learning. *J. exp. Psychol.*, 1949, 39, 575-579.
- SMITH, M. H. Spread of effect is the spurious result of non-random response tendencies. *J. exp. Psychol.*, 1949, 39, 355-368.
- THORNDIKE, E. L. A proof of the law of effect. *Science*, 1933, 77, 173-175.
- THORNDIKE, E. L. An experimental study of rewards. *Teach. Coll. Contr. Educ.*, 1933, No. 580.
- TOLMAN, E. C. Connectionism; wants, interests, and attitudes. *Charact. Pers.*, 1936, 4, 245-253.

(Accepted for publication February 4, 1957)

Psychological Monographs: General and Applied

The In-Basket Test¹NORMAN FREDERIKSEN, D. R. SAUNDERS, AND BARBARA WAND
Educational Testing Service

I. INTRODUCTION

THE Officer Education Research Laboratory, Air Force Personnel and Training Research Center, has concerned itself for some time with the problems of evaluation of the effects of instruction in Air University. In connection with this work, Educational Testing Service was asked to study the desired outcomes of training in the Command and Staff School (a part of the Air University, formerly known as the Field Officer Course) and to develop prototype methods to determine how well these objectives are being achieved. This report describes an attempt to develop such prototype methods. Although the research was directed toward evaluation of the curriculum at the Command and Staff

School and not toward assessment of individual officers, the discussion has distinct implications for the development of instruments to evaluate individual performance.

The problems of assessment in areas demanding a high level of performance present a challenge to those interested in new techniques of measurement. There is a clear need for instruments which will measure such complex skills as the ability to organize discrete pieces of information, to discover the problems implicit in a situation, to anticipate events which may arise because of such problems, and to arrive at decisions based on a large number of considerations. These and other skills are continually demanded of administrative officers in key positions.

At this level of functioning, tests of in-

¹ This research was supported in whole or in part by the United States Air Force under Contract No. 33(600)-5833, monitored by Officer Education Research Laboratory, Air Force Personnel and Training Research Center, Maxwell Air Force Base, Alabama.

The project was conceived by the Human Resources Research Institute when it was under the command of Major General Carroll. This organization has since become the Officer Education Research Laboratory and a unit of the Air Force Personnel and Training Research Laboratory, Air Research and Development Command. Dr. Don-Samuel M. Goodman and his successor, Dr. Donald J. Malcolm, have acted as contract monitors; they and their staff have been very helpful on numerous occasions. The project could not have been conducted without the active collaboration of Colonel Walker and Colonel Adams, successive Commanders of the Command and Staff School, ACSC, Air University, and their Curriculum Planning Board. Lieutenant Colonel Sheeks and

Lieutenant Colonel Wall, members of this Board, served effectively as our liaison with CSS. Colonel Ritchey, Director of General Courses, acted as liaison between ACSC and ETS for the first phase of the project. Dr. T. F. Staton, Director of Educational Assistance in ACSC, has been particularly helpful in the course of the study.

Acknowledgment is due a number of ETS staff members who worked on the project at various times and in various capacities. During the first phase of the project Dr. Warren Findley was Principal Investigator and Dr. Paul Diede- rich, Dr. Paul Freeman, Dr. Harold Gulliksen, Dr. William G. Mollenkopf, and Mr. Charles Allen contributed in various ways to the study. Dr. Gulliksen and Dr. Mollenkopf aided in the development of the in-basket problems. Miss Henrietta Gallagher provided technical help in the analysis of the data and Mrs. Marjorie Tulinski provided assistance in scoring. Dr. Irving Lorge served as a consultant to the project.



tellectual ability bear a lower relation to performance than they do to performance on tasks of a less complex nature, partly because selection on the basis of intelligence has already taken place, and partly because administrative responsibilities appear to demand additional skills.

Several earlier attempts to devise situational tests which would evaluate performance at this level have been reported in the literature. For instance, use has been made in Great Britain of such tests in selecting men for responsible posts in the Civil Service, for selecting industrial executives (2, 3), and by the War Office Selection Board for evaluating officer potential (4). In Great Britain, however, the tendency has been to use group problem-solving techniques more frequently than written tests. In Australia, Lafitte (6, p. 107) reports the development of a written device which, in its conception, is not unlike the test described in this monograph.

Considerable imagination and versatility have been demonstrated in the work on these new techniques. At the same time, the proponents (7) of these innovations have frequently expressed their scepticism of the value of objective, scorable, psychological tests. Thus, it is often the case that in attempting to solve the problems of evaluation of performance in high-level jobs the test-maker is considered to be facing the choice of constructing objective, reliably scored, but relatively insensitive instruments, or developing a more sensitive measure which resists attempts to use it reliably and objectively.

The instrument which is described in this report is the result of an attempt to devise a sensitive measure which may at the same time be objectively and reliably scored, and proceeds from a faith that progress toward both goals of sensitivity and of objectivity may be made in one operation.

This instrument, which has been called the In-Basket Test, is a situational test presented in written form and group administered. The briefing on the nature of the problems and the presentation of the problems are carried out in such a way that the information available to the candidate is the same for all candidates. The test allows a great freedom in response. The problems are presented in such a

way that it is up to the candidate first to discover the problem and only then to organize an attack. Although the In-Basket Test was designed to represent the situation faced by the Field Officer in the Air Force, material suitable to other areas of experience may readily be adapted to this form.

It is hoped that the description of the steps taken in developing this instrument and of the problems encountered along the way, as well as the recommendations for further improvement of the test, may be helpful to those who are working on instruments of this kind.

The first phase of the research involved a careful study of the curriculum and of the objectives of the course in order to determine quite specifically what aspects of their work the officers were expected to perform more effectively as a consequence of the training.² Students in the Command and Staff School (CSS) are mostly of the rank of major and lieutenant colonel and have been specially selected for training to fit them for greater administrative responsibilities as field grade officers. This training includes courses with such titles as Organization, Management, Personnel, Intelligence, Operations, and Logistics. The main source of information about the objectives of the instruction was a series of interviews with instructors and school officials. The instructors were asked to state what on-the-job activities they would expect graduates to handle more expertly as a consequence of the instruction. An attempt was made to get the instructors to avoid generalities and to describe observable behaviors which would indicate whether or not a student had attained the desired objective.

² This phase of the study was carried out under the direction of Dr. Warren G. Findley (1).

Altogether more than 500 such statements were collected. In order to put this large number of behavioral descriptions of desired outcomes into a more orderly system, the next step was to classify the statements into categories which would correspond to psychologically meaningful functions.

Twelve categories make up the system of classification which was evolved. In six of these categories behaviors are primarily *individual*; that is, they are behaviors which could be exhibited by a person all alone. The other six categories involve behaviors that are primarily *interactive*; that is, they involve a relationship with other people. Four of the individual categories were selected as the primary focus of the evaluation instrument to be developed.

The categories may best be defined by first giving some examples of statements and then the term that has been chosen to represent the statements in that category.

Below are some selected statements of objectives which fall into one of the categories:

Carefully follows established supply procedure under normal circumstances.

Deals only with the civilian personnel officer in all matters of civilian employment.

Refrains from making inappropriate requests of units on an Air Force base.

These statements assert that the graduate of the Command and Staff School is likely to make *efficient use of routines*, employing actions appropriate to the scope of his assignment, and using channels and "SOP" (Standard Operating Procedure) to advantage.

Here are some examples of another type of behavior:

Readily plans for and makes changes in pro-

cedures which are consequences of the introduction of new research developments.

Changes the organizational pattern of a base without hesitation when it seems desirable.

Uses the B-29 bomber not only for high-level bombing, but also for tactical support at low level, for delivering napalm bombs, and for other unconventional uses.

These statements are alike in demanding *flexibility*, adaptability, and willingness to introduce change.

Here are some additional examples:

Can comprehend the effect on AF activities of possible political-economic events such as the closing of highways and railways into Berlin.

Anticipates and makes projections of his plans many months in advance.

Prepares alternative operational plans, based on different contingencies.

These statements assert that the graduate of the Command and Staff School is relatively likely to show *foresight*, that is, to anticipate the feel of future situations. This involves anticipating possible as well as probable consequences and providing for contingencies.

Here are some statements which fall into still another category:

In selecting weapons to achieve a stated objective, considers the various available weapons first from the viewpoint of efficiency, and second from the viewpoint of economy.

In deciding upon or advising concerning a particular decision to be made, includes cost as an important factor.

Plans for making appropriate use of various weather conditions in combat operations.

These statements assert that the graduate of the Command and Staff School is likely to *evaluate data effectively*. This will involve judgment as to what data to include as pertinent and what to exclude as irrelevant to a solution.

All four of these categories of behavior are individual behaviors. The other two

categories of individual behaviors are *knowledge* and *effective guidance of the decision-making function* of the unit.

Since the interactive behaviors were expected to require relatively unwieldy evaluation methods, priority was given to developing measures of the individual types of behavior. Also, some of the curriculum divisions, such as Communications and Electronics, and Judge Advocate General, were thought not to justify immediate or extensive evaluation efforts, in view of the small proportion of time devoted to them in the instruction. Thus, four curriculum divisions—namely, Management, Personnel, Operations, and Logistics—and four major types of individual outcomes—namely, efficient use of routines, flexibility, foresight, and effective evaluation of data—were selected to receive the main emphasis in the development of evaluation methods.

II. GENERAL DESCRIPTION OF THE IN-BASKET TEST

Purposes of the Evaluation Instruments

A measurement device was planned which would provide a separate score for each of the four selected functional categories of behavior, and which would support curriculum development by yielding information pertinent to the related curriculum divisions. Thus, the In-Basket Test resulted from an attempt to develop measuring instruments which would permit an evaluation of the extent to which students profited from the aspects of the instruction which aimed at improving their ability to use Standard Operating Procedure (SOP), increasing their flexibility, improving their foresight, and increasing their ability to evaluate data effectively. The intent was not merely to discover if the students had mastered textbook knowledge about flexibility, for instance, but rather to find out whether

or not the graduates of the course exhibited in their own behavior on the job the characteristics which were being sought. From one point of view, a *criterion* measure was being developed as a means for evaluating the effectiveness of Air Force administrative officers.

Description of the In-Basket Materials

The materials which have been developed are called collectively the In-Basket Test. A large amount of the daily work of an administrator centers around the contents of his in-basket. The In-Basket Test consists in putting a candidate into a realistic situation which calls upon him to deal appropriately with such material as an Air Force officer might find in his in-basket.

The form of the test which was tried out at Maxwell Air Force Base involved eight hours of testing, during which each candidate was required to play four roles in succession. In one two-hour period the candidate played the role of a Commanding Officer of a hypothetical Composition Wing, in another the Wing's Director of Materiel (D/M), and in other periods Director of Personnel (D/P) and Director of Operations (D/O). In each of these roles he was given an in-basket containing incoming letters, memoranda, staff studies, letters prepared for his signature by subordinates, and other similar material. He was given suitable forms on which to write answers, and his directions were to go to work as though he were actually on the job.

A common, and often valid, objection to situational tests is that the only reasonable response to a situation is, "It depends." Examinees feel that it is unfair to be presented with a complex situation in four lines and then be expected to describe the best action. An attempt was made to overcome this difficulty in the

case of the In-Basket Test by presenting a sufficient amount of information to make it unreasonable to say that the only correct answer is "It depends." This was done in two ways. First, a considerable amount of background information about the hypothetical composite wing was provided for study by the candidates prior to the time when they actually took the test. This information included such items as a statement of the mission of the wing, a short history of the wing, an organization chart, a table showing the strength of the wing, a roster of the key officers of the wing, maps showing the location of Pine City Air Force Base, and maps indicating the landing strips and buildings of the Base. Second, the situation was so arranged that an adjutant supposedly had studied the contents of the in-basket and had added from the files appropriate documents to explain the letters or memoranda and to provide needed background information. In making a decision in a real-life situation one never has *all* the relevant information. In this respect the test situation was probably not greatly different from a situation in real life.

The development of the problems which became parts of the In-Basket Test was accomplished through close cooperation with officers of the Air Force, and involved writing letters, memoranda, and staff studies which were like those which actually might be found in the in-basket of an Air Force officer. This phase of the work is described in greater detail in a later section.

When the candidate appeared to take the test, he had already been given an opportunity to study the background materials. His instructions were roughly as follows:

Today you are asked to take the role of Director of Personnel of the 71st Composite Wing.

The previous Director of Personnel, Lt. Col. Hart, was killed in an auto accident, and you have been assigned to take his place. A manila envelope on your desk contains the materials which have collected in Hart's in-basket together with additional material placed there by the Adjutant for your guidance. Your job is to read your mail and take appropriate action as though you are actually on the job. Write the appropriate notes, memos, letters, or directives. Take as much action as you can with the information available to you. You are limited in that no more information can be obtained during the next few hours and you can communicate only in writing.

In his Director of Personnel in-basket, the candidate found such items as (a) a letter for the policy file from Col. Goodfellow, the Wing Commander, which states that it will continue to be the policy of the wing to pay constant attention to the problem of morale; (b) a letter from the Personnel Officer regarding the conduct of Airman Third Class Joe Doakes, who is a personnel problem, asking that Doakes be given a talking-to; (c) a note from Col. Goodfellow requesting that an appropriate policy statement with respect to hardship and bad conduct discharges be drafted; this note is backed up by a memorandum from the Legal Officer describing how punishment had been escaped by an airman through the hardship discharge, and another document showing that the former Director of Personnel has concurred in the Legal Officer's recommendation; and (d) a memorandum from the Civilian Personnel Officer regarding a visit by a delegation of civilian employees who have a grievance. These are only a few of the problems found in the in-basket.

There is no one-to-one relationship between problems and documents presented to the candidate. Some of the documents are included merely as statements of policy—as additional background information. Sometimes a problem is represented by a combination of two or three documents which are not physically

together. In other instances the real problem may not be the one stated by the writer of a memorandum, but rather one which is implicit in what is stated.

III. HOW THE IN-BASKET TEST WAS DEVELOPED

Development of the Setting

Logically, the first step was to develop some concept of a basic situation or situations in which the problems would be set. After some of the problems had been substantially prepared, it was decided to place all of the problems in the setting of an imaginary "71st Composite Wing." Since all four roles are placed in the same imaginary wing, a minimum of confusion was created for the subjects who are required to play each role in turn; also, an opportunity was created for the examinees to carry over information learned in one role to subsequent roles. By placing the imaginary wing in Maine, an opportunity was created to use problems dependent on adverse weather conditions and, by making the wing a "composite" wing, opportunity was created to use problems involving several types of aircraft.

The actual process of developing the situation went hand in hand with the development of specific problems that would fit the situation, until a considerable body of factual detail about the 71st Composite Wing was built up. Forms were developed, tables of strength and organization were drawn up, statements of mission and policy and records of performance were created, all in order to provide the background information needed to set and solve the problems. It is believed that the background information provided was broad enough to make it reasonable for candidates to take action on a majority of the problems.

Use of Essays in Constructing Problems

The first approach that was systematically used to create problems for the In-Basket Test involved study of a series of essays written by the students in the CSS soon after they arrived at the school. Students were asked to describe some problem in the Air Force—how it arose, the factors bearing on its solution, what had been done to solve it, and what remained to be done to complete its solution. The essay was written as if to brief a successor to the job when there was no opportunity for personal briefing. Approximately half of the materials developed for the July 1953 tryout of the in-basket procedure were derived from ideas in these essays.

Use of Interviews in Constructing Problems

It was also considered desirable to tap some source of problems remote from Air University in order to eliminate any possible bias toward, or away from, problems familiar to students in the course. Accordingly, arrangements were made to visit McGuire Air Force Base, New Jersey, for the purpose of developing additional problems for the test. Interviews were held with about a dozen of the officers in the Headquarters of the Air Defense Wing and in the Groups directly under this wing located at McGuire Air Force Base. The aim of the interview was always to isolate problems which looked as if they would meet the requirements of the in-basket situation. An attempt was made to block out each problem in sufficient detail so that actual writing of the necessary materials could proceed. This included a discussion of the general nature of each memorandum, letter, or other document that it would be necessary to prepare.

Steps in Problem Preparation

The following steps typically were applied in preparing a problem for the In-Basket Test from materials obtained from the sources discussed above.

1. A suitable point was selected in the development of the problem at which its presentation would require a minimum amount of reading.
2. It was determined whether or not the problem could be related to one of the four roles in the 71st Composite Wing (Commander, D/M, D/O, or D/P).
3. A method was determined for presenting the problem in this role. This method sometimes involved dividing the problem up into several memoranda, so that the problem would be found only by noticing the relationships between the memoranda. Or in other cases it was handled by preparing a poor solution for presentation along with the problem, since correspondence that is passed forward without recommendations is properly returned as "incomplete staff action."
4. Materials that were necessary to present the problem were written. An effort was made to ensure that these materials were stated briefly and clearly, so that the understanding of the problem would not depend too much on reading comprehension. Indications were given that the materials had been seen by everyone who should have seen them before they got into the in-basket (unless the purpose of the problem was to test the student's recognition of incomplete coordination).
5. Other supporting documents were written that might be needed to complete the picture of the problem or to provide information that would be needed for the intended solution.
6. The problem materials were then reviewed by someone familiar with Air Force terminology, to ensure that the language was correct and unambiguous and that the materials did not unintentionally deviate from normal command and staff procedures.
7. Several persons were asked, as an informal tryout, to indicate the action they would take on the problem; this helped to ensure that most subjects would not be sidetracked and that the differences in performance would be of the intended kind, in response to the intended problem.

Development of the Procedure for Administration of the Test

In the main tryout of the test in July 1953, the following steps were taken in introducing the testing situation and in attempting to build up motivation:

1. The proctor reads a statement outlining the ground rules of the test situation and "explaining" how the subject had gotten into it. For example, the following instructions were for the role of D/P (which was not the first role played).

a. Today you are asked to take the role of the Director of Personnel of the 71st Composite Wing, Pine City AFB, Maine. The previous D/P, Lt. Col. Charles Z. Hart, was killed four days ago in an auto accident, and you have been assigned to Pine City AFB to take his place. *Today's date is Tuesday, 6 July 1954.* The manila envelope on your desk contains the materials which have collected in Hart's in-basket, plus additional material placed there by the Adjutant for your guidance. . . .

b. Your job is similar to the previous ones. You are merely to read your mail and take appropriate action as though you were actually on the job. Write the appropriate notes, memos, letters, or directives. Take as much action as you can with the information which is available to you. As before, you are limited in that (1) no more information can be obtained during the next two hours, and (2) you can communicate only in writing.

c. Please write legibly. Place the appropriate file number on each sheet, and sign your own name. I have additional copies of Form 71W3³ if you need them.

d. When you have finished, place all of the materials back into the envelope, including the materials you have written. Place a paper clip on all the papers you have written on in order to separate them from papers you have not written on. Leave any unused Forms 71W3 outside of your envelope.

³ The subjects were asked to use Form 71W3 in writing all their messages. This is a form, designed for use in the test, which is similar to "buck slip" forms in use in the Air Force, but not identical to any (see Fig. 1).

- lope. Write your name and course code number on the outside of the envelope.
- e. As you know, the findings of this test will have no bearing on your grades, but may be very helpful in the development of evaluation methods. Therefore, you are requested to cooperate by doing your best.
 - f. Go ahead.

Development of the Scoring Procedure

Once the problems and a procedure for their administration had been developed, the major remaining problem was the development of an adequate system for scoring the test on the various categories of functional behavior.

Having administered the series of problems in the tryout, samples of answers were drawn for each problem recognized for scoring purposes. Three samples were drawn, which will be referred to as Samples I, II, and III. Sample I consisted of the responses of one student randomly drawn from each of the 39 cubicle-groups used in administering the In-Basket Test. Sample II was similarly drawn from the cases remaining after constructing Sample I. Sample III was drawn from the cases remaining after constructing Sample II, by randomly selecting *three* officers from each cubicle. It was considered desirable to stratify the samples on the basis of their cubicle assignment in order to eliminate any possible effects of communication between subjects during the course of the testing. It is known that there was some communication in some of the cubicles, but it is believed that this was largely confined to procedural matters of the test administration.

The responses of the students to each problem in Sample I were sorted into categories representing the various types of response to the problem. For example, *concurring in a recommendation* might be one type of response (although the concurrence might be indicated in a great variety of ways); *refusing to concur* might be another type; and *referring the problem to higher authority* might be a third type of response. Such categories were further combined in an effort to arrive at a relatively small number of types of response that were clearly distinct from one another and easily definable. Between five and fifteen types of response were identified for each of the various problems, based on a study of the responses in Sample I.

In setting up these types of response it was necessary to ignore a great deal of information and to abstract from each response those aspects of the response which were relevant to the functional category of behavior (e.g., foresight or

flexibility) which the problem was supposed to measure. Even striking differences from student to student in style or tone of the communication were ignored in setting up the types of response. This abstracting of a limited aspect of the behavior necessarily meant loss of a great deal of information but presumably had the advantage of making the scores reflect only the variable being measured.

The types-of-response lists, obtained from study of Sample I responses, were tested by using them to classify the responses from Sample II. It was found necessary, in order to make the system work for Sample II, to redefine or broaden the concepts of some of the types of response. On the whole, the types of response obtained from the first sample were found to be adequate.

On the basis of the study of Sample I and II responses, some problems were found to be unsuitable because of too great a uniformity in the answers or because of too great a scatter in types of response. Such problems were eliminated, so far as further scoring is concerned. In other instances it was found advisable to score the problem for a different functional category of behavior than it was originally designed to measure.

Following the study of Samples I and II, final lists of types of response were prepared for each remaining problem. These were submitted to two panels of expert judges in the Air Force. One panel consisted of twelve students in the Air War College (AWC). The other panel consisted of twelve staff members from Air Command and Staff College (AC & SC) (of which CSS is a part). Each panel was chosen to include three officers particularly qualified to judge the problem materials in each of the four roles of the In-Basket Test.

Each subpanel of three officers was asked to consider the problems in the in-basket for the appropriate role and to assign scoring values to each type of response listed for each problem being scored. The subpanels from AWC worked independently from the subpanels from AC & SC, but each subpanel worked as a team.

The judges were first asked to consider the in-basket as a whole, and to assign a rank-order of importance to the problems being scored. The level of priority given to each problem by the judges was later used in determining the scoring weight assigned for failing to respond to the problem. The agreement between the rankings of two subpanels of judges was invariably almost perfect. The judges were then asked to consider each scorable problem separately, and to assign scoring values on a five-point scale to each type of response, such that each point on the scale was used at least once for each problem.

The amount of agreement between the two panels of judges on the value to be assigned to the types of response was quite variable, ranging

TABLE 1

CORRELATIONS BETWEEN SCORES ASSIGNED BY TWO PANELS OF EXPERTS AND SCORING WEIGHTS FINALLY ASSIGNED TO TYPES OF RESPONSE FOR EACH IN-BASKET PROBLEM^a

Problem	Correlations Between Scores Assigned by Two Panels of Experts (AWC and ACSC)	Correlations Between a Com- posite of the Scores of Both Panels and Final Scoring Weights	Correlations Between Scores Assigned by AWC and Final Scoring Weights	Correlations Between Scores Assigned by ACSC and Final Scoring Weights
C059	.64	.70	—	—
C213	.42	.87	—	—
C229	.50	.62	—	—
C446	.60	.84	—	—
C496	.38	.96	—	—
C525	.74	.92	—	—
C650	.82	.92	—	—
C705	.66	.86	—	—
C747	.47	.99	—	—
C983	.58	.81	—	—
Mo87		.80	—	—
M215	.77	.91	—	—
M278	.51	.78	—	—
M325	.57	.96	—	—
M534	.79	.85	—	—
M552	.66	.88	—	—
M916	.59	.82	.00	1.00
M941	.64	—	—	—
P058	.00	.95	—	.77
P107	.35	—	.62	—
P139 ^b	.26	—	—	.59
P247	.00	—	.13	—
P406	— .20	.88	—	— .20
P461	.49	—	.94	—
P526	— .27	.64	.70	.27
P607	.43	—	—	.28
	.01	—	.87	—
O116		—	—	—
O290	.07	.68	—	—
O301	.72	.78	—	—
O500a ^b	.65	—	.42	.58
O500b	— .31	—	—	—
O568	.22	.86	—	—
O705	.40	.78	—	.91
O786	.54	.93	— .28	—
O946	.66	—	—	—
	— .28	—	—	—

^a In cases where the agreement between the Air Force panels falls below .35, separate correlations with the AWC and ACSC panels are reported.

^b No final scoring weights assigned.

from good to poor. Occasionally even negative correlations were observed. The first column of Table 1 presents the correlation for each problem.

The test-makers, of course, had their own ideas about what the scoring weights should be. In the case of some problems there was substantial agreement between all three groups—the AC & SC panel, the

AWC panel, and the test-makers. In other instances the two military panels were in poor agreement; here it was usually found that the test-makers agreed substantially with one of the panels. In only two of the problems was the lack of agreement so great as to suggest that the problem be dropped.

The decisions made in determining the

final scoring values were generally arrived at by the following process:

1. A careful examination was made of the weights suggested by the Air Force panels, in order to determine which of the original functional classes of behavior was apparently being given greatest weight in determining their judgments. These determinations were compared with the original notion of the functional class of behavior for which the problem had been written. A decision was then made as to the functional class to which the final scoring of the problem would be assigned. By this time, a number of problems had become assigned to another of the curriculum objectives, namely, Guidance of Decision-Making.

2. The various types of response were ordered by the score developers in terms of their knowledge of the functional category of behavior determined in step 1. In doing so, they attempted to avoid being influenced by the detailed rankings made by the Air Force experts. This procedure was designed to ensure that the final scoring weights would reflect as nearly as possible a single functional category of behavior, but would at the same time correlate as highly as possible with the judgments of the Air Force panels of judges.

An indication of the success of this procedure is provided in Table 1, in which are presented the actual correlations between the composite (over-all average) Air Force judgments and the final scoring weights. (In cases where the agreement between the Air Force panels as shown in column one of Table 1 falls below .35, separate correlations with the AWC and AC & SC panels are reported.)

Scoring a paper, then, consisted in reading a response to a problem, comparing the response with the listed types of response for that problem, deciding to which type the particular response belonged, and then assigning the numerical value which corresponded to that type.

Sample III of the responses selected from the CSS data were not looked at until a final scoring summary sheet for that problem had been prepared and the scoring manual had been written. The results obtained in the scoring and anal-

ysis of Sample III are presented in the next section of this report.

IV. RESULTS OF THE COMMAND AND STAFF SCHOOL TRYOUT

The "in-basket" materials were tried out in July 1953 by administering them to the entire Class 53-B of the Command and Staff School. The administration required four two-hour sessions, which were distributed at intervals of about three days through the second and third weeks of the course. Each student in the course played in turn the roles of Commander (CO), Director of Materiel (D/M), Director of Personnel (D/P), and Director of Operation (D/O) of the 71st Composite Wing. The purpose of the July 1953 tryout of the materials was to provide data which could be used to determine some of the operational and statistical characteristics of the In-Basket Test, as well as to develop the scoring method. These characteristics of the test will be discussed under four major headings: Scoring Reliability, Over-all Reliability, Validity, and Attitudes Toward the Test.

Scoring Reliability

Reliability of scoring may be assessed by comparing the scores assigned by one scorer with scores assigned by a second scorer, and is expressed here in terms of the product-moment correlation coefficient between the scores assigned by two scorers of the same set of responses.

The scoring reliability may be expected to be a function of the amount and quality of training that the scorers have had. Appropriate training would ordinarily involve direct experience in scoring a large number of answers and in discussing the scoring of these answers with other more experienced scorers. Relevant experience with the job situation which the In-Basket Test is designed to simulate would be very helpful in learning to score the test.

TABLE 2
CORRELATIONS BETWEEN SCORERS A, B, AND C
FOR SINGLE DIRECTOR OF MATERIEL
PROBLEMS AND FOR D/M
TOTAL SCORE

Problem	Scorers A & B N = 50	Scorers B & C N = 20	Scorers A & C N = 20
M087	.91	.78	.80
M215	.85	— ^a	— ^a
M278	.88	.87	.85
M325	.54	.77	.55
M534	.69	.76	.69
M552	.63	.66	.47
M916	.77	.48	.51
M941	.89	.89	.94
D/M Total Score	.83	.69	.68

^a Scores not available for scorer C.

Two sets of scoring reliability data have been obtained. The first set of data is for three independent scorers all of whom can be regarded as having had at least some relevant previous experience. Two of these scorers (A and B) were primarily responsible for the development of the scoring procedures and manual and were presumably the best scorers available. Scorer C had some contact with the military situation portrayed, as well as with day-to-day administrative problems in his own work, but had not been intimately associated with the development of the scoring procedures. The correlations reported in Table 2 are for the first 50 cases of Sample III D/M problems for scorers A and B and the first 20 cases for scorer C.

The second set of data pertaining to scoring reliability is for two independent scorers, B and D. Scorer D worked directly from the final form of the scoring manual without much special training and without benefit of prior experience with the military situation. These data indicate how reliably the In-Basket Test may be scored by someone working mostly from the printed instructions with a minimum of special training. Table 3

presents the resulting correlations for each problem, for the total score for each role, and also for the grand total score. Even under these conditions scoring reliability is reasonably satisfactory. It is clear from these data, considered as a whole, that the In-Basket Test can be scored with a reasonably high degree of reliability.

Over-all Reliability

The problem of determining the over-all reliability of the In-Basket Test is somewhat unusual. The rationale of the test calls for the measurement of a series of distinct, conceptually different dimensions—namely, the functional categories of behavior identified in the study of curriculum objectives. Therefore, one aspect of reliability which was of interest was the degree to which problems designed to measure the same dimension of behavior did, in fact, measure the same thing. At the same time, the structure of the test provides four distinct roles, which are the separately timed portions of the total test. This permitted a test of the reliability of the total score of the In-Basket Test under the assumption that the four roles are four equivalent forms of the test. Thus, a two-way breakdown of the problems is possible; and in designing the testing battery an effort was made to provide problems that could be sorted into the cells in such a two-way breakdown. It should be noted here that all values indicating over-all reliability are based on the scores assigned by one scorer, scorer D.

The most direct evidence bearing on the initial hypotheses that underlay the preparation of the items and their assignment to functional categories is provided by the intercorrelations of the items intended to measure a particular category of behavior. One would not expect the intercorrelations of single items to be high. Among items which make up a test or subtest, however,

TABLE 3

CORRELATIONS BETWEEN SCORERS B AND D FOR SINGLE IN-BASKET PROBLEMS, ROLE TOTALS, AND GRAND TOTAL FOR COMMAND AND STAFF SCHOOL SAMPLE III

Commanding Officer Problems			Director of Materiel Problems		
Problem	N = 112	Correlation	Problem	N = 110	Correlation
C059		.92	M087		.71
C213		.75	M215		.81
C229		.87	M278		.81
C446		.89	M325		.65
C496		.81	M534		.84
C525		.83	M552		.62
C650		.93	M916		.69
C705		.81	M941		.90
C747		.76			
C983		.88			
Total Score on CO Problems		.91	Total Score on D/M Problems		.74
Director of Personnel Problems			Director of Operations Problems		
Problem	N = 102	Correlation	Problem	N = 106	Correlation
P058		.83	O116		.83
P107		.73	O290		.91
P247		.79	O301		.64
P406		.75	O500		.83
P461		.80	O568		.86
P526		.80	O705		.78
P607		.70	O786		.79
			O946		.83
Total Score on D/P Problems		.80	Total Score on D/O Problems		.86
Total of All Problems $r = .90$					

generally positive correlations would be required in order to show that the items are homogeneous in the sense that they all tend to measure the same ability. Such homogeneity is a sufficient condition for high reliability.

Intercorrelations of items intended to measure particular functional categories are presented in Tables 4, 5, 6, 7, and 8. It is evident that there is not a high degree of homogeneity among the problems. In the cases of Categories II, III, IV, and V (the four categories originally established as the objectives for measurement), the correlations are predominantly positive; in the case of the problems placed in Category VI, Guidance of Decision-Making, even this cannot be said. Category III (Flexibility) evidently shows the highest amount of homogeneity, and Category VI (Guidance of Decision-Making), the lowest amount. Category VI was not originally included among the functional categories to be represented in the test; the large representation of this category in the test resulted from the reclassification of problems because of the opinions of the Air Force committees. All problems in this category

were originally written to represent other functional categories.

None of the categories shows enough homogeneity to warrant computation of a separate reliability coefficient. If such a coefficient were computed for the reliability of the Category III total score, it is guessed that it would be equal to about .40 or .50.

It will be noticed that in some of the tables of intercorrelations, one or two problems account for most of the negative values. In Table 4, for example, problem O290 accounts for both of the negative correlations, and in Table 8 problem P461 is consistently negative. As in ordinary test construction procedures, the homogeneity and hence the reliability of the score could be improved by eliminating these items which detract from the total test. As it is, seven per cent of the correlations reach the one per cent level of significance.

TABLE 4

INTERCORRELATIONS OF PROBLEMS PLACED IN
FUNCTIONAL CATEGORY II
(USE OF ROUTINES)

Problem	C446	M916	P607	O290
C446		10	09	-05
M916	10		08	10
P607	09	08		-05
O290	-05	10	-05	

TABLE 5

INTERCORRELATIONS OF PROBLEMS PLACED IN
FUNCTIONAL CATEGORY III
(FLEXIBILITY)

Problem	Co59	M215	P107	P247	O568
Co59		01	31**	03	01
M215	01		04	12	42**
P107	31**	04		-06	07
P247	03	12	-06		11
O568	01	42**	07	11	

** Significant at the .01 level.

TABLE 6

INTERCORRELATIONS OF PROBLEMS PLACED IN FUNCTIONAL CATEGORY IV
(FORESIGHT)

Problem	C747	M534	M941	P526	O116	O500	O705
C747		-02	06	18	11	-06	17
M534	-02		00	17	31**	15	-13
M941	06	00		02	-08	-08	13
P526	18	17	02		09	05	12
O116	11	31**	-08	09		28**	-03
O500	-06	15	-08	05	28**		20
O705	17	-13	13	12	-03	20	

** Significant at the .01 level.

TABLE 7

INTERCORRELATIONS OF PROBLEMS PLACED IN FUNCTIONAL CATEGORY V
(EVALUATION OF DATA)

Problem	C650	Mo87	M278	P406	O301	O786	O946
C650		14	-10	-17	03	00	-02
Mo87			-30**	-02	22*	17	-01
M278	-10	-30**		05	-14	13	28**
P406	-17	-02	05		08	27**	11
O301	03	22*	-14	08		05	06
O786	00	17	13	27**	05		04
O946	-02	-01	28**	11	06	04	

* Significant at the .05 level.

** Significant at the .01 level.

TABLE 8

INTERCORRELATIONS OF PROBLEMS PLACED IN FUNCTIONAL CATEGORY VI
(GUIDANCE OF DECISION-MAKING)

Problem	C213	C229	C496	C525	C705	C983	M325	M552	Pos8	P461
C213										
C229										
C496	09		22*	05	-05	14	-07	13	10	-07
C525	22*	25*	25*	-02	-02	08	-23*	-06	03	-17
C705	05	-02	-19	-19	-10	13	-03	-06	-01	-11
C983	-05	-02	-10	06	06	09	-11	05	11	-06
M325	14	08	13	09	00	00	-02	-11	06	-00
M552	-07	-23*	-03	-11	-02	-25*	-25*	20	14	-01
Pos8	13	-06	-06	05	-11	20	03	11	02	-03
P461	10	03	-01	11	06	14	-03	-01	-09	-01
	-07	-17	-11	-06	-00	-01				-09

* Significant at the .05 level.

The lack of evidence of a high degree of homogeneity should not be surprising in view of the subjective nature of the categories. An empirical approach to the problem of classification into functional categories may prove to be more valid than the judgmental approach which was used.

The next problem to consider involves the question of whether or not the In-Basket Test as a whole may provide a meaningful total score. Evidence on this point is provided by the intercorrelations of the total scores for the four roles, and the correlations of these with the grand total score. This information is presented in Table 9. The numbers in Table 9 vary widely, with the correlations between role total scores ranging from .06 to .45. The three smallest correlations all involve the relations between total score for D/P and the other scores; none of these three correlations is statistically significant. The other three correlations, which do not involve the D/P score, are all significant beyond the .05 level, using a one-tailed test of significance. Thus, aside from the D/P score, there appears to be some homogeneity in the materials covered by the different roles. (Whether the D/P score is a reliable measure of something else, or is merely an unreliable measure of the same function measured by the other three roles, cannot be answered by the data in Table 9.)

The information contained in Table 9 provides a basis for estimating the reliability of the total In-Basket Test. This works out to be .50 for the eight-hour battery, and .42 for the six-hour battery obtained by omitting the D/P problems from the total test. The former estimate of .50 is probably too high because the assumption of equivalent parts of the test is not satisfied (8).

These values for reliability, while they are of definite statistical significance, are not of sufficient magnitude to justify the interpretation of an individual's scores

TABLE 9

INTERCORRELATIONS OF THE ROLE SCORES AND CORRELATIONS OF ROLE SCORES WITH GRAND TOTAL SCORE ON THE IN-BASKET TEST FOR COMMAND AND STAFF SCHOOL SAMPLE III^a

	CO	D/M	D/P	D/O	Grand ^b Total
CO					68
D/M		.24	.14	.19	69
D/P	.24		.15	.45	50
D/O	.14	.15	.06		67
Grand Total ^b	68	69	50	67	

^a $N=92$.

^b Correlations of Role Scores with the Grand Total Score are, of course, spuriously high since the Grand Total Score is based on the sum of the four Role Scores. The expected "chance" value of these correlations is about .50.

on the In-Basket Test. Applying the Spearman-Brown prophecy formula, it appears that approximately 24 hours of testing time would be required in order to achieve a reliability of the order of .75 or more with material like that in the present In-Basket Test. However, it must be remembered that the present test is the first such instrument constructed, and the present tryout is a pretest of this first experimental form. It seems probable that higher reliability could be obtained by selection of the best problems, by development of better problems to replace those found to be unsatisfactory, and by improvement of the scoring categories and methods.

The obtained values for reliability for the present form are sufficiently high to justify comparing groups, for investigating correlations between scores and various criteria, etc. The original purposes for which the In-Basket Test was devised were all of this type, and it may therefore be concluded that, insofar as reliability is concerned, the original objectives of the test can be met using the materials made available in this research.

Another important objective of this

research was the differentiation of several different desired outcomes of instruction in the CSS. It was hoped that scores could be obtained from testing that would make it possible to determine the relative success of the instruction in achieving these different desired outcomes. Is there any evidence in the present data that would suggest that different meaningful outcomes can be differentiated? If there is, it may compensate for the failure of the a priori categories of behavior to be clearly reflected in the empirical findings.

In order to make possible the discovery of one or more categories of behavior, a complete table of intercorrelations among all the individual problems included in all four roles was computed. This table was inspected in an effort to isolate clusters of at least five problems for which all the intercorrelations were relatively large and consistent in direction, in the sense that if a problem correlated negatively with *one* of a number of positively correlated problems it must correlate negatively with all of them.

Four clusters meeting this requirement were found; four of them are presented in Tables 10, 11, 12, and 13. The fourth cluster, it will be noted, consists of six variables. Several general observations may be made concerning these tables.

First, the four clusters presented involve, all told, fifteen different variables, which is almost one half of the individual problems available for inclusion in clusters. If clusters of four instead of five variables were sought, many of the remaining problems could be brought into clusters. This result seems to justify the inference that most of the individual problems do possess some degree of reliability over and above mere scoring reliability. The obtained estimates of over-all reliability could thus be improved by appropriate regrouping of problems.

A second general observation is that not all of the problems are positively intercorrelated with the rest of the problems in their clusters. In particular, problem P461, which occurs in three of the clusters, has consistently negative correlations with the other problems in all three

TABLE 10

CLUSTER 1

Problem	M215	O568	O500	O786	P461
M215	—				
O568	42	—			
O500	27	19	—		
O786	24	15	17	—	
P461	-16	-14	-20	-12	—

TABLE 11

CLUSTER 2

Problem	M215	O500	O568	C496	P461
M215	—				
O500	27	—			
O568	42	19	—		
C496	22	24	14	—	
P461	-16	-20	-14	-11	—

TABLE 12

CLUSTER 3

Problem	O500	M534	O116	P461	O301
O500	—				
M534	15	—			
O116	28	31	—		
P461	-20	-23	-11	—	
O301	27	14	12	-18	—

TABLE 13

CLUSTER 4

Problem	P526	P607	C213	C983	M552	Po58
P526	—					
P607	17	—				
C213	25	28	—			
C983	24	16	14	—		
M552	14	20	13	20	—	
Po58	17	09	10	14	11	—

clusters. A related observation is that the problems scored as measures of functional Category II (Use of Routines) tend to have more negative correlations with other problems than do problems scored for the other functional classes of behavior. In particular, negative correlations are observed between problems scored for Class II (Use of Routines) and Class III (Flexibility). Both of these observations suggest that the desired outcomes of instruction in the Command and Staff School are not merely nonhomogeneous, but are to some extent contradictory and

conflicting in the manner of their interrelation within the student population. *The possibility must seriously be considered that achievement towards one objective will simultaneously represent movement away from some other objective.* The achievement of a proper balance between objectives thus becomes an even more difficult problem.

A third general observation based on Tables 10 through 13 is that it might be possible, by some method such as factor analysis, to obtain an empirical set of classes of functional behavior which would describe the domain of the problems better than the a priori system of classification. Such a factor analysis would need to be applied to the complete table of intercorrelations and would be more likely to provide useful results if an even larger number of problems were available for simultaneous analysis. In view of the generally small magnitude of the problem intercorrelations, it would be appropriate to conduct the factor analysis according to some rigorous method, such as the Lawley Maximum Likelihood Method, to ensure that the bounds of statistical significance were not being exceeded by the number and nature of the factors extracted.

It is of possible interest to offer an interpretation of some of the clusters represented in Tables 10 through 13. Each cluster is likely to represent a group of problems that would turn out to have loadings on the same factor if a factor analysis were performed.

Let us first consider Table 10. All of the problems in this cluster seem to imply *a desire on the part of the officer to make judgments suited to the merits of the individual cases, without regard to standard procedures which may or may not indicate the same conclusion.* Thus, in problem M215, both the best and poorest answers involve acceptance of the endorsement to an inspection report, but the best answer also indicates the intention of conducting a special inspection of the ammunition storage facilities. This problem best represents the cluster, in view of its relatively high correlations with the others. In problem O568, both good and poor answers involve following a training regulation. The good answer wants to establish the "intent" of the regulation, and then to provide an SOP "interpreting" it as it applies to the present situation, whereas the poor answer merely calls for "strict" compliance. In problem O500, both good and poor answers call for examination of a three-hour turn-around time, which is too long. Here, the good answer calls for an initial meeting to discuss the problem; whereas the poor answer

plunges immediately into a detailed study of the problem, which may be unnecessary. In problem P461, which indicates that hardship discharges have been given to airmen in cases where a bad conduct discharge would have been more appropriate, decisions that move toward the preparation of a policy statement on the matter generally are regarded as better than those moving toward consideration of the individual case; and we observe, accordingly, that scores on this problem have a negative correlation with the rest of the cluster.

The cluster presented in Table 11 has four problems in common with the one just discussed, and can be interpreted along the same psychological lines. Most of the problems in the clusters represented by Tables 11 and 12 were intended to measure the Flexibility category of behavior. They may perhaps be thought of as representing a more homogeneous, "purified" set of flexibility items.

Let us now consider Table 12. All of the solutions correlating with this cluster seem to imply *a willingness to take definite action on problems that clearly call for definite action, even on problems that clearly call for prolonged consideration.* Thus in problem M534, which best represents this cluster, the scoring is entirely in terms of the amount of concrete action taken in the direction of grounding possibly defective F94 aircraft. Problem O116 is very similar, and is scored in terms of the amount of concrete action facilitating an investigation of the fouling of C119 carburetors. In problem O500, one of the two best answers is to set up immediately a maintenance training program, while the poorest answer is to do nothing. In problem P461, any definite decision regarding a policy statement on hardship and bad conduct discharges is given a poorer score than an answer delaying the decision; even taking no action at all is given a positive scoring weight. Accordingly, this problem is negatively correlated with the remainder of the cluster. Again, in problem O301, the definite decision to include various factors in a solution earns higher scores, and taking no action earns relatively lower scores. All four of the problems positively correlated with this cluster were originally planned as measures of the Foresight category.

A third example may be taken from Table 13 which contains the intercorrelations of six problems. All of the problems seem to demonstrate *a working knowledge of the organization and appropriate functions of various offices, including the office currently held by the subject taking the test.* The exact nature of the action taken in each problem varies. Thus, in problems P058 and P607, the basic problem needs to be referred to an appropriate person, at either lower or higher echelon. In problems C213 and C983, more in-

formation needs to be gathered, and the recommendations of the basic problem need to be modified. In problem P526, there is enough information to justify immediate concurrence. In problem M552, advice should be given, but personal charge of the project should not be taken. Most of the problems in this cluster came to be regarded as measures of the "Guidance of Decision-Making" class of behavior.

In general, the evidence presented supports the results of the armchair analysis but suggests that an improved basis of classification of school objectives could be found which would permit the development of reasonably homogeneous subtests of sufficient reliability for individual measurement. Further improvement might come from empirical refinement of the scoring weights attached to the types of response. The present test materials seem to provide an adequate basis for group comparisons and correlational studies, by utilizing the group of problems within the experimental in-baskets for which there is some evidence of content reliability. The reliability of the In-Basket Test in its present pretest form is satisfactory for certain limited uses (those involving comparisons of groups of reasonable size), and with further development the reliability might become sufficiently high to justify use of the test in individual selection or placement.

Validity

The term validity may be thought of in several senses. The simplest concept of validity is "predictive" validity; predictive validity may be measured by the correlation between scores on the test and some measure of success in the activity the test is supposed to predict. The In-Basket Test is not primarily intended to forecast success. It was intended rather as a measure of achievement which would be useful in the evaluation of instruction.

How can an achievement test be validated? It might be argued that an

TABLE 14
CORRELATIONS OF ROLE TOTAL SCORES AND
GRAND TOTAL SCORE WITH ACE TOTAL
SCORE, AND CSS FINAL GRADE FOR
CSS SAMPLE III^a

Variables	ACE Total Score	CSS Final Grade
Total Score CO Problems	.24	.14
Total Score D/M Problems	-.01	.10
Total Score D/P Problems	.12	.12
Total Score D/O Problems	.25	.02
Grand Total Score	.25	.15

^a $N = 92$.

achievement test does not need to be validated; insofar as the test directly represents the skills it is intended to measure, the test is "self-validating." Probably no one would quarrel about such a definition of validity for a test of (say) long division. In the present instance, however, there might be more disagreement about whether or not the test does represent the skills taught; the objectives of the instruction in the Command and Staff School are not quite as neatly defined by a set of test items as are the objectives of an arithmetic course.

In the long run, the validation of a test depends upon building up a wealth of experience which is consistent with a particular definition of a test. We tend to accept a test as valid to the extent that over a period of time it is found to correlate positively with the measures which we feel the test should correlate with and not with things we feel it should not correlate with. Gulliksen (5) has called such a concept of test validity "intrinsic validity."

Correlations between In-Basket Test scores and four other kinds of data have been computed. None of these other vari-

TABLE 15
BREADTH OF EXPERIENCE AS RELATED TO ROLE TOTAL SCORE ON THE IN-BASKET TEST

Primary Air Force Specialty Code	Narrow Experience										Total Est. Mean		
	n	CO Est. Mean	Est. SD	n	D/M Est. Mean	Est. SD	n	D/O Est. Mean	Est. SD	n		D/P Est. Mean	Est. SD
Personnel Administration Maint. & Supply Operations	16	0.31	4.69	14	1.14	2.41	16	-2.56	4.47	14	2.14	4.05	0.26
	19	-0.47	4.45	19	2.16	4.33	18	-1.78	4.83	16	4.41	4.91	1.09
	8	-4.12	6.20	8	-0.25	5.44	8	-2.25	5.92	7	2.86	1.57	-0.94
	34	-2.50	4.79	35	1.31	3.59	32	-1.88	3.64	31	3.81	3.42	0.18
	$F=2.05$ Not significant			$F=3.69$.05>p>.01			$F=0.10$ Not significant			$F=0.79$ Not significant			
Primary Air Force Specialty Code	Broader Experience										Total Est. Mean		
	n	CO Est. Mean	Est. SD	n	D/M Est. Mean	Est. SD	n	D/O Est. Mean	Est. SD	n		D/P Est. Mean	Est. SD
Personnel Administration Maint. & Supply Operations	2	3.00	2.83	2	-1.50	0.71	2	0.50	6.36	2	3.00	1.41	1.25
	8	0.88	2.80	4	4.50	6.32	5	-1.40	2.70	5	3.40	3.36	1.82
	8	1.00	6.87	8	4.00	3.21	7	-0.71	5.71	8	2.88	4.00	1.79
	3	-2.00	3.61	9	1.44	2.30	9	-2.56	4.10	9	3.56	4.25	0.11
	$F=0.45$ Not significant			$F=0.34$ Not significant			$F=0.36$ Not significant			$F=0.07$ Not significant			

ables is a "criterion." The correlational results are presented merely as a contribution to our understanding of what is and what is not measured by the In-Basket Test.

The four kinds of data are as follows: (a) scores on the American Council on Education Psychological Examination (the "ACE"); (b) CSS course grades; (c) statements about type of Air Force experience; and (d) statements about breadth of Air Force experience.

Table 14 presents the correlations of the first two kinds of data with the four role total scores and over-all total score on the In-Basket Test. The over-all total score correlates .25 with ACE total score, which is clearly significant from a statistical viewpoint. Such a correlation with a measure of mental ability is to be expected in a test that measures any sort of intellectual function.

Since the ACE total score and the CSS grades are correlated only to the extent of -.02 in our sample, we may regard the CSS grade as being completely independent of ACE score. The correlation of the in-basket total score with CSS grades was

found to be .15. While this correlation is on the borderline of statistical significance when evaluated with a one-tailed test against the null hypothesis, certainly it does not represent a relationship of much practical importance. It cannot be determined from the available data whether the low correlation observed here is due to relatively low reliability for the CSS grades (as well as for the present form of the In-Basket Test), or to lack of any strong relationship between the two measures. In view of the observations noted above which suggest that some of the objectives of the course may prove to be diametrically opposed to other objectives of the course, we perhaps should not expect a very high correlation. It must be remembered also that the In-Basket Test was administered early in the course, not as a final examination.

Table 15 presents information involving the third and fourth types of data mentioned above. This table is divided into two parts, according to breadth of administrative experience. The larger group of students, represented at the top of the table, are considered to have had

relatively narrow administrative experience, because their Primary Air Force Specialty Code (PAFSC) at the time of testing fell in the same general area as their stated area of greatest experience. The remainder of the group, at the bottom of the table, are considered to have had relatively broad experience, because their PAFSC at the time of testing did *not* fall within the same general area as their greatest experience. Within these major groupings, the cases have been divided according to the stated area of greatest experience, the areas having been chosen so as to correspond as closely as possible with the roles of the In-Basket Test. For each subgroup the table shows the mean and standard deviation for each of the four roles of the In-Basket Test.

The data in Table 15 were analyzed using analysis of variance, and the F values which resulted are shown. Comparing the score variance within these subgroups, we find that all but two of the F values are less than unity, and one out of eight values reaches significance at the .05 level of confidence. In view of the number of tests made, even this result does not warrant further consideration. It should be noted that the great heterogeneity of variances *within* the groups based on area of greatest experience theoretically makes the F test invalid. However, the effect produced by this imperfection in the data would be in the direction of producing apparent significance when it should not exist. Since the results do not appear significant anyway, the interpretation seems to be surely justified that area of greatest administrative experience has little if anything to do with performance on the In-Basket Test. This is an important conclusion, because it means that there is no necessity, for reasons other than face validity, to construct a special form of the In-

Basket Test to examine officers in various areas of specialization.

The other comparison permitted by the data of Table 15 is between those individuals having broad and those having narrow administrative experience. These comparisons were originally suggested by a close study of the results for the D/M problems, where it was found that those with broad experience had D/M total scores significantly better than those with narrow experience, at the .05 level of confidence. As soon as the scores for the CO problems were available, and the data were subjected to the same kind of treatment, it was found that the officers with broad experience did better, and that this difference was significant at the .01 level of confidence. However, when the relative performance of these groups on the D/P and D/O problems was assessed, no significant differences were found.

None of this information settles the question of the validity of the In-Basket Test. The findings do represent a beginning in the development of our understanding of the nature of the test. No conclusive evidence that the test is valid can be offered. In the final analysis, validity is confidence in a test which is generally borne out by numerous observations about the test over a period of time.

Attitudes Toward the In-Basket Test

Following the administration of the In-Basket Test, a memorandum was mailed to the CSS students. This memorandum provided the students with additional information concerning the purpose of the project in which they had cooperated.

A list of eight proposals for possible future applications of the new technique was included in the memorandum. These uses were stated, as follows:

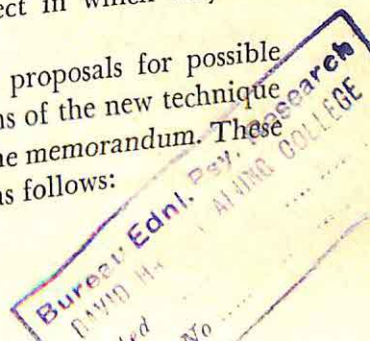


TABLE 16

JUDGMENTS OF 111 CSS STUDENTS WITH REGARD TO EIGHT PROPOSALS FOR FUTURE
USE OF THE "IN-BASKET" TEST MATERIALS

Proposed Use of Test ^a	Judgment			
	"Suitable"	"Possibly Suitable" "Of Doubtful Suitability"	"Not Suitable"	No Comment
1. Basis for officer evaluation at command and staff levels	22	10	64	4
2. Basis for selection of the superior officer for special assignments	28	9	58	5
3. Basis for group comparison	66	7	18	9
4. Absolute standard of performance in command and staff level jobs	6	5	82	7
5. Indicator of group strengths and weaknesses	65	8	21	6
6. Realistic work sample for preparing officer to face his new job	51	6	41	2
7. Instructional materials	59	9	24	8
8. Further technique of final evaluation in school such as CSS	44	14	36	6

Note.—Each cell shows the percentage of the total group expressing a judgment.
^a For exact statement of proposed uses see page 24.

1. To provide a basis for uniform, impartial evaluation of officer effectiveness at the command and staff level.

2. To provide a basis for selecting superior officers for certain types of assignments.

3. To provide a basis for comparing one group of officers with another group which may have had different job or educational experience, or both.

4. To provide an absolute standard of performance in command and staff level jobs.

5. To provide information to a school such as CSS concerning the general strengths and weaknesses of a group of students.

6. To provide students in a school such as CSS with a realistic sample of the work they will later meet in command and staff jobs of the Air Force.

7. To provide instructional problems for which most of the possible alternatives are known in advance.

8. To provide a further technique of final evaluation of individual students in a school such as CSS.

The students were invited to evaluate the test materials in the light of the eight proposals, and to include suggestions for the improvement of the materials if they saw fit. The questionnaire was worded as follows:

It is requested that each student indicate his answer to the following two questions:

(a) Considering the test materials as you saw them, and answering from your personal point of view, do you consider them adequate for the purposes stated above in 2-c? (Insofar as your answer may be negative, please be specific.)

(b) Do you have any comments or suggestions for improving the test materials, for the purposes stated above in 2-c? (Again, if you do, please be specific. You may omit any comment that applies only to a single instance in the test materials.)

Of the 500 or so copies of the memorandum that were distributed, 335 were filled out and returned. The majority of the respondents chose to discuss the materials generally, rather than to relate their comments to any specific proposed application as it was outlined in the memorandum.

However, 111 of the students did relate their comments to specific proposals. Their opinions are represented in Table 16. Within this group of 111 the tendency seems to have been to approve of applications of the in-basket technique to groups, but to object to its use in the evaluation of individuals. These comments, it will be

remembered, applied to the actual test materials seen, not to the general test idea or to some improved test which might be developed.

Of the other 224 respondents, 42 felt the materials were adequate, but had no suggestions to offer in the light of any of the proposals. The remaining 182 evaluated the materials in a general way, making no attempt to relate their comments to the specific proposals. The responses exhibited a great range of opinion, varying from profound skepticism of the test to enthusiastic approval.

The most telling criticisms pointed to a need for revision of the briefing materials. Some involved requests for amplification of the information contained in the briefing, such as more reference material on the policy of the wing, fuller organizational and functional charts, more information regarding predecessors' duties, and some reference to the handling of classified material prior to placing in the in-basket.

Perhaps even more vital to the ultimate success of the test were suggestions that the briefing should ensure an appropriate "set" on the part of the student. Some felt that the importance of the project should have been more heavily stressed if full cooperation was to be obtained. Others felt that a fuller and more frank explanation of the purpose of the test would have helped in obtaining this cooperation.

A large number of students expressed what might be interpreted as hostility toward the testing procedure due to what they felt was an "artificial" restriction on normal modes of communication. Many suggested that memo writing is an unimportant tool of the good staff officer, and that his skill is exhibited, rather, in personal contact with his staff.

Some of the respondents suggested that this attitude toward the test created by the restriction on communication might be somewhat allayed by more adequate preparation in the initial briefing. It was suggested that the examiners display in the briefing their awareness of these limitations. Some students felt that specific reference to means by which the student could indicate the content of hypothetical phone calls, staff meetings, or informal conversations would be helpful.

A large number of students objected to the problems on the basis that they were "unrealistic," although the problems have been based upon real events in the Air Force. Others felt that the high proportion of problems exhibiting poor staff action gave an unrealistic picture of the base as a whole.

Some criticisms were directed toward other aspects of the test which deviated from actual experience, such as failing to provide an opportunity for group discussion and cooperative

solution of the problems. Some felt that insufficient background material precluded any basis for wise action. Those who acted, they felt, would be making superficial judgments. Others were convinced that the short time allowance militated against considered judgments.

V. RECOMMENDATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The present In-Basket Test represents a first attempt to develop an instrument for use in evaluation of education at the very high level of complexity which exists in the Command and Staff School at Air University. It is a prototype of an instrument which might become valuable not only for training evaluation but also for individual selection, placement, and guidance. Now that the pretest of this measuring device has been completed, recommendations can be made about the use of the In-Basket Test in its present or an equivalent form, and suggestions can be offered which have to do with further development of the test.

Recommendations for Use of the Test in Its Present Form

1. The In-Basket Test may justifiably be used in making comparisons between mean scores of groups of examinees. This is the use for which the test was originally intended. Thus a group of CSS students tested at the beginning of the course might be compared with a group tested at the end of the course. Or a group of CSS graduates might be compared with a group of officers who are similar to the graduates except that they have not been students in CSS. If the groups to be compared are reasonably large, at least some of the part scores for roles and for functional categories could be used in such comparisons as well as total score. Analysis of variance is the preferred method of analysis if suitable measures can be found to provide a basis for controlling ability factors in these comparisons.

2. The In-Basket Test may also be used as instructional material. Administration of homogeneous parts of the test at appropriate points in the course, followed by critiques of the performance of the students would seem to be an excellent instructional device. This is a use of the test which was not planned, but which would be

warranted to the extent that instructors found the problems suitable for their objectives.

3. The frequencies with which the various types of response occur in an administration of the In-Basket Test may enable instructors to gain greater understanding of the results of their instructional efforts as they apply to specific performances of students. For example, instructors might find that in view of the large proportion of students making one of the low-rated types of response, many students must have failed to generalize some point in the instruction. On the other hand, they might be pleased to find that a new method of presentation had increased, from one class to the next, the proportion of students giving a "good" type of response to several similar problems. Such detailed analysis of performance on an examination provides maximum guidance to the instructor who wishes to improve the quality of his teaching.

4. The test might be used as a method for assessing a group of entering students in order to find out their general level of capability in the skills measured by the test. Thus the level or amount of instruction might be varied in accordance with the ability of a particular entering class, or the students might be "sectioned" into two or three levels of ability in order that the level and amount of training would be more suited to their needs.

5. The test should *not*, in its present form, be used as a basis for individual evaluation, either for selection, placement, guidance, or measurement of course achievement.

Suggestions for Further Research

The pretest form of the In-Basket Test represents a novel approach to a measurement problem, and little recorded experience was available to guide the development of the test and scoring procedure. Consequently the tryout has revealed a number of ways in which, it is felt, the test and testing procedure could be improved. The following suggestions concern improvement of the test with respect to the breadth of information reflected in its scores, and the reliability of measurement.

1. If one reads a sample of responses to the In-Basket Test, he will probably get the impression that much more information about the examinees is contained in the responses than is reflected in the scores reported. The scoring

procedure requires the scorer to decide which of several types of response a particular response resembles, and this involves throwing away a good deal of information. The first recommendation, therefore, has to do with *extending the breadth of information obtained from the test by suitable revision and extension of the procedure for evaluating the responses.*

One revision or extension of the scoring system proposed here is to give the scorer greater freedom in awarding or subtracting points for exceptionally good or poor performance with respect to categories of behavior other than the one for which the problem is primarily being scored. For example, in a response to a problem primarily intended to measure foresight, the examinee might write something which indicates outstanding ability to make effective use of standard procedures. Thus he would be given credit for this response by tallying a point in the "SOP" column of his score sheet.

Similarly, evaluations might be made of characteristics that the test was not originally intended to measure and that perhaps were not revealed in the content analysis of outcome statements. For example, various degrees of cooperativeness seem to be revealed in the writing of many of the examinees. Some students go beyond the call of duty in offering their services in their memoranda, while other students never indicate in their writing a spontaneous willingness to help someone else work out a problem. If these statements reflect a genuine attitude of cooperativeness, it would seem desirable to work out a procedure for recording the instances as they are found in the scoring process. Another characteristic which at least occasionally seems to be revealed is the attitude associated with sympathy or harshness in dealing with subordinates. The In-Basket Test, in other words, may be thought of as a projective test which reveals a great deal of as a projective test which reveals a great deal of more about the personality of the writer than is revealed by the objective scoring method described in its Appendix. Developmental work leading to modification of the test and evaluation procedure in an attempt to capitalize on the breadth of information which seems to be contained in the responses seems to be justified.

2. The most obvious fault of the In-Basket Test in its present form is its low reliability. While low reliability might to some extent be compensated for by breadth of information, greater reliability is certainly desirable and probably could be attained without necessarily decreasing breadth of information. The following suggestions have to do with *increasing the reliability of the In-Basket Test scores.*

The development of the scoring procedure involved making a content analysis of the responses to a particular problem and then asking military

experts to evaluate the resulting types of behavior. One possible reason for low reliability might be too great an influence of military as opposed to psychological judgment in the ordering of these types of responses. High reliability implies homogeneity of the items of a test with respect to the variable being measured. The procedure used in this research may have introduced more impurity than was necessary or desirable.

A second suggestion pertaining to the improvement of reliability is that the homogeneity of the subtests be increased by suitable analytical procedures. The major categories of behavior which correspond to the most important subtest scores (Foresight, Flexibility, etc.) were obtained by classifying the statements of outcomes of instruction given by the CSS faculty. An analysis of the responses to all the test problems by factor analysis (or some analogous method) might lead to the discovery of more valid bases for grouping items into subtest scores. The cluster analysis already performed tends to support the argument that homogeneous subtests would result from such a procedure.

Finally, improvement of reliability should result from more stringent selection and revision of items.

In addition to the specific suggestions made above, it should be emphasized that if the test is to function as an adequate evaluation device it must frequently be used and revised by the instructional staff. The involvement of instructors, rather than outside staff, in the continued development and revision of the test will be more likely to ensure that the results of the testing will be *used* and that the content of the test is appropriate.

VI. SUMMARY

This report describes the preparation of test materials which were designed to aid in the evaluation of instruction in the Command and Staff School of Air University. A major purpose of this course is to increase the administrative proficiency of field grade Air Force officers. Test materials for evaluation of instruction were considered desirable in order to support improvement in curriculum planning and instructional procedures.

The test which was developed is a situational test which requires the examinee to play four roles: Commanding Officer, Director of Materiel, Director of Personnel, and Director of Operations of a fictitious Composite Wing. The examinee is provided with background information about the wing and the Air Force base where it is located, suitable material from the "files," and the contents of an in-basket which is appropriate to each role (hence the name, In-Basket Test). The in-basket contains letters, memoranda, and other documents which embody problems aimed at eliciting behavior relevant to the objectives of instruction in the course. The examinee's task is to write the letters and memoranda which he would write if he were actually on the job, to sign (or not to sign) letters prepared for him, or to take other suitable action. Scoring of these products is intended to provide information which would be useful to the Air University in evaluating the instruction in the Command and Staff School. Full instructions for administering and scoring the test are provided in its Appendices.

This report also presents the results of a tryout of the test. The purpose of this tryout was to investigate some of the statistical properties of the test, student reaction to the test and the adequacy of the instructions and procedures for administration. An entire class in the Command and Staff School took the test in the third week of the course.

The results indicate that the In-Basket Test can be scored with a reasonably high degree of reliability, but that the present form of the test is low in content reliability. There is evidence, however, that the content reliability could be improved by reassigning scoring weights to the re-

range of information obtained from the test, and for increasing its reliability through reassessment of scoring weights,

improvement in the homogeneity of the subtests, revision of items, and changes in the scoring procedure.

APPENDIX

Several sample problems in the form of facsimile letters taken from the in-basket of the Commanding Officer are reproduced below. The letters which appear in memo form, with the notation "Form 71W3" at the lower left, were each typed on a form specially designed for this test. It is similar to "buck slip" forms used in the Air Force, but not identical to any (see Fig. 1).

The first problem is presented by a memorandum from the Adjutant requesting permission to correct his log of classified documents in an unauthorized manner. This problem is intended to measure Efficient Use of Routine.

The second problem involves a directive to all pilots, prepared by the Director of Operations for the Commanding Officer's signature as a result of a complaint that pilots have been buzzing the campus. The best solutions recognize that the Director of Operation's proposed action is unimaginative and that a specific positive program for prevention of such violations is required. The problem is intended to measure Foresight.

The third sample problem involves a complaint from a used car dealer that two airmen have passed a bad check. The candidate should realize that the investigation of the Legal Officer was inadequate and that the Commanding Officer should insist on a more thorough investigation before signing a letter such as has been prepared for him to sign. The problem is intended to measure Guidance of Decision-Making.

In the fourth problem the Wing Chaplain presents a letter, for the Commanding Officer's signature, requesting the Ministerial Association to develop increased recreational facilities. The best action is one which recognizes that a much broader study of the recreation situation is needed. The problem is intended to measure Guidance of Decision-Making.

PROBLEM 1

To: Commandant

FROM: Adjutant

FOR: Action

FILE: C446 DATE: 3 July 195X

SUBJECT: Correction of the Log of Classified Documents.

REMARKS:

1. The Administrative Inspector, HQ, 99th AF, discovered a discrepancy between our Log of Classified Documents and the File of Classified Documents.

2. The missing Document is Secret Letter MV 163-092-21, dated 16 March 195X, from HQ AMC.

3. The report of the Administrative Inspector contains no mention of the specific missing document, but merely mentions a discrepancy between the Log and the File of Classified Documents which must be corrected.

4. The document described in Paragraph 2 was burned by the undersigned after appropriate action on the letter had been taken. There were unfortunately no witnesses; the undersigned was new at his job and did not realize fully the importance of the action. There is no question that the letter was burned.

5. Permission is requested to correct the Log of Classified Documents by drawing up a new Log which omits the listing of the document described in Paragraph 2, and to destroy the present Log of Classified Documents.

(signed) JOHN T. SMITH
Major, USAF
Adjutant

Form 71W3

PROBLEM 2

First Letter

EASTERN UNIVERSITY
CAROL, MAINE

Office of the President

June 25, 195X

Colonel Hiram W. Goodfellow
Commanding Officer
Pine City Air Force Base
Pine City, Maine

Dear Sir:

I am writing you about a situation that has been causing our students, instructors, and administrative officers increasing concern. Our University has a number of returned fliers in the classes. They are utilizing planes at Pine City AFB for flying, I suppose partly for amusement

and partly to earn the appropriate number of flying hours.

They are, however, abusing this privilege tremendously to the detriment of the University. A great deal of flying is done over the campus, buzzing not only the University buildings, but also buzzing students when they are walking along relatively open stretches on the campus. The noise of the diving planes interferes with the lectures and recitations; furthermore, other pilots in the class seem to take it as a challenge and vow they will come closer than their buddy did just then. So far no one has been seriously injured or killed; however, it would seem that this would not be unlikely if present antics are continued. The University is doing everything in its power to cooperate with the military and to assist in the present emergency. We trust that you will be able to prevent a further occurrence of such unnecessary annoyances as I have referred to.

Yours truly,
(signed) MILTON F. JONES
President
Eastern University

MFJ: sbm

Second Letter

To: Director of Operations

FROM: Commandant

FOR: Action

FILE: C747 DATE: 27 June 195x

SUBJECT: Flying Violations

REMARKS: Please look into this and suggest appropriate action.

Form 71W3

(initialed) H. W. G.

Third Letter

To: Commandant

FROM: Director of Operations

FOR: Signature

FILE: C747 DATE: 2 July 195x

SUBJECT: Flying Violations

REMARKS: I have looked into the facts alleged in President Jones' letter. He has not exaggerated the situation.

(signed) W. T. THOMPSON
Lt. Col., USAF
Director of Operations

Form 71W3

Fourth Letter

To: Air Operations Inspector

FROM: Commandant

ATTN: All pilots

FOR: Action

FILE: C747 DATE: 3 July 195x

SUBJECT: Flying Violations

REMARKS:

1. It has been brought to the attention of this Command that there have been repeated violations of paragraph 13, AFR 60-16, Minimum Altitude of Flight, in the vicinity of the Eastern University Campus, Carol, Maine.

2. Since this Command has not been informed of these violations under the provisions of paragraph 49, AFR 60-16, Authorized Deviations, it is assumed that these incidents are recognized by the pilots concerned as being unauthorized.

3. No useful military purpose is served by such conduct.

4. The attention of pilots violating cited regulations is invited to paragraph 9c, AFR 36-57, Causes for Suspension, Serious Willful Violation of Flying Regulations.

(signed) HIRAM W. GOODFELLOW
Colonel, USAF
Commander

Form 71W3

PROBLEM 3

First Letter

W. DOWNE EAST

THE SMILING ESKIMO!

Pine Plaza West
Pine City, Maine

PERSONAL

June 28, 195x

Colonel H. W. Goodfellow
Commanding Officer
Pine City Air Force Base
Pine City, Maine

Dear Colonel:

Two weeks ago, two soldiers stationed at your base bought a used car from me for \$400. The soldiers were Sgt. Myron Q. Jones and Sgt. Herbert K. Snyder. They each gave me a check for \$203.98, drawn on the National Shawmut Bank of Boston, with the story that these were

their pay checks for May. The checks have been returned marked "Depositor Unknown." They asked me not to report that they were out of uniform, although this was also true.

The car has since been sold to Sgt. Homer H. Dinger, according to the State Registrar of Motor Vehicles. Sergeants Jones and Snyder have been sent overseas, according to reports given me by soldiers still living in their quarters.

I wish you would do something about this. Either Sergeant Dinger should return the car to me, or pay me for it. Jones and Snyder should be brought back to stand trial for bad check passing.

Sincerely yours,
(signed) W. DOWNE EAST

WDE: pac

Second Letter

To: Legal Officer

FROM: Commandant

FOR: Action

FILE C-496 DATE: 1 July 195X

REMARKS: Please draft an appropriate reply to this letter, for my signature.

(initialed) H. W. G.

Form 71W3

Third Letter

71st COMPOSITE WING
PINE CITY, MAINE

C496

3 July 195X

Mr. W. Downe East
Pine Plaza West
Pine City, Maine

Dear Mr. East:

Your letter of 28 June, criticizing the actions of three of the airmen of this command, has been referred to my legal officer for consideration. We are sorry to learn of your experiences with them.

However, we are powerless to be of any assistance to you in this matter. Sgt. Dinger's title to the car is evidently clear, since the State Registrar of Motor Vehicles has accepted it. He cannot be expected to pay for the same car twice.

Although we did send a Sgt. Myron Q. Jones and a Sgt. Herbert K. Snyder overseas last week, there is no proof that they were not impersonated. Sgt. Dinger tells us that he conducted all his negotiations with them by long distance telephone the day after they left here for overseas. He wired them their money to Seattle, and they

mailed him his title. While this is a highly unusual way of doing business, the story checks in every detail except for positive identification of Jones and Snyder.

Sincerely yours,

(for signature of) HIRAM W. GOODFELLOW
Colonel, USAF
Commander

PROBLEM 4

First Letter

To: Commandant

FROM: Chaplain

FOR: Signature

FILE: C525 DATE: 2 July 195X

SUBJECT: Community Facilities for Recreation

REMARKS:

1. In view of the projected expansion of this base, I feel that recreational facilities in Pine City will soon become a matter of critical concern.

2. I have prepared the attached letter for your signature. I sincerely hope that you will be able and willing to speak to the Ministerial Association on this matter. Of course, if you so desire, I should be glad to do so in your place, but I sincerely believe that your personal attention to this matter would insure the success of this project.

(signed) S. Y. SOLE
Major, USAF
Wing Chaplain

Form 71W3

Second Letter

71st COMPOSITE WING
PINE CITY, MAINE

3 July 195X

C525

The Rev. Mr. Simon T. Pureheart, Chairman
Pine City Ministerial Association
Old North Church
Pine City, Maine

Dear Reverend Pureheart:

You have probably seen the announcement from Washington that the facilities of Pine City Air Force Base are to be markedly expanded.

We have in the past enjoyed generally harmonious relationships with the townspeople. There have been numerous instances of activities

that have been set up in Pine City to provide good, wholesome recreation and diversion for our men during off-duty hours. However, the doubling of the base complement which will occur over the next six months poses new problems.

I am therefore writing to you as head of the Ministerial Association to ask that your group consider what plans the churches might make

so as to help provide increased facilities for off-duty activities. I shall be glad to speak before your group if you should desire to invite me to do so.

Sincerely yours,

(for signature of) HIRAM W. GOODFELLOW
Colonel, USAF
Commander

REFERENCES

1. FINDLEY, W. G., FREDERIKSEN, N. O., & SAUNDERS, D. R. An analysis of the objectives of an executive-level educational program. Maxwell Air Force Base, Alabama: ARDC, Hum. Resour. Res. Inst. Tech. Res. Rep. No. 22, Jan., 1954 (Contract AF 33 (600)-5833).
2. FRASER, J. M. An experiment with group methods in the selection of trainees for senior management positions. *Occup. Psychol.*, 1946, 20, 63-67.
3. FRASER, J. M. New-type selection boards in industry. *Occup. Psychol.*, 1947, 21, 170-178.
4. GARFORTH, G. I. DE LA P. War Office Selection Boards. *Occup. Psychol.*, 1945, 19, 97-108.
5. GULLIKSEN, H. Intrinsic validity. *Amer. Psychologist*, 1950, 5, 511-517.
6. LAFITTE, P. Melbourne Test 90. *Aust. J. Psychol.*, Monogr. Suppl., 1954, No. 1.
7. RODGER, A. Selection for management. *Occup. Psychol.*, 1947, 21, 203-206.
8. TUCKER, L. R. Part-whole correlations and test reliability. Research Memorandum 51-13 (dittoed). Princeton, N.J.: Educational Testing Service, 1951.

(Accepted for publication February 11, 1957)

Psychological Monographs: General and Applied

Forced Choice and Other Methods for Evaluating Professional Health Personnel

SIDNEY H. NEWMAN, MARGARET A. HOWELL
*United States Public Health Service*AND FRANK J. HARRIS
Operations Research Office, The Johns Hopkins University¹

INTRODUCTION

THIS study was undertaken to compare the forced choice technique with other methods for evaluating the performance of commissioned professional health personnel in the United States Public Health Service. It is a part of the officer selection and evaluation program described by Newman (7).

The Public Health Service, the major Federal organization responsible for the health of the nation, employs approximately 16,000 Civil Service personnel and 3,000 commissioned officers. The commissioned officer component of the Service is composed of carefully selected personnel in various scientific specialties and in the health professions of medicine, dentistry, nursing, sanitary engineering, pharmacy, veterinary medicine, dietetics, and physical therapy. These officers hold clinical, research, public health, and administrative positions in Service hospitals, outpatient clinics, research centers, regional offices, other governmental agencies, and foreign countries.

The performance evaluation of commissioned officers is accomplished through periodic efficiency reporting. In an effort to improve the Service's performance-rating system, an investigation of methods for evaluating job performance was undertaken in 1949. A review of the literature in the performance-rating field revealed that no data had been reported then, or at present, on the kinds of highly trained scientific and professional personnel employed in the varied and specialized work areas in the Service.

A consideration of various performance evaluation methods led to the con-

clusion that the forced choice technique developed by the Department of the Army appeared most promising for use in the Public Health Service setting. While investigations of the forced choice technique have been based on populations which are quite different from Public Health Service professional personnel, reports by Sisson (10), Witsell (12), and the Adjutant General's Office, Department of the Army (13, 14) seemed to indicate the usefulness of the technique in a commissioned personnel system.

An Experimental Efficiency Report, incorporating forced choice items and other evaluation materials not in use in the Service in 1949, was designed and distributed to the supervisors of active-duty officers. The effectiveness of the forced choice technique is here compared with that of the more conventional evaluation methods included in the Experimental Report. It is anticipated that the results of this study will contribute to the literature on performance evaluation and, in particular, to that of forced choice methodology. The findings may also be of special interest to local or state health departments, hospitals, research institutions, or other organizations employing professional personnel engaged in public health work, medical care, or research relevant to the problems of

¹ Formerly with the United States Public Health Service.

health and disease. In addition, this study calls attention to some of the problems implicit in the evaluation of relatively small numbers of employees engaged in a variety of specialized professional activities.

PROBLEMS

An investigation of performance-evaluation methods was undertaken to answer the following questions:

1. Is the forced choice technique an effective method for measuring the performance of professional health personnel in the Public Health Service?
2. How does the forced choice technique compare in validity and reliability with more conventional methods of performance evaluation?
3. How do factors such as the characteristic evaluated in criterion ratings, the administrative level of the supervisor completing efficiency reports, and the grade of the ratee affect the validity of efficiency reporting?
4. What combination of efficiency-reporting methods optimally predicts the performance of personnel in the various professional and occupational fields of the Service?

In addition to the major problems of the study, it was possible to compare the validity of the Experimental Report with that of the Officer's Progress Report (the efficiency report in use in the Service in 1949) which had been completed under operational rather than experimental conditions.

MATERIALS

Criteria

The criteria of performance within the Public Health Service consisted of 20-point graphic rating scales used for the evaluation of each of the following factors: Work Performance, Administra-

tive Ability, Personality (Personal Qualifications), and Over-all Value to the Service. Instructions for each scale requested raters to compare the ratee with a typical group of personnel having similar duties and responsibilities. A rating of 1 was used to designate the least effective ratees, and a rating of 20 was used to designate the most effective.

Experimental Efficiency Report

This Report was divided into the following four sections, samples of which may be seen in the Appendix.

Section I—Forced Choice. This part consisted of 50 tetrads adapted from items developed by the Department of the Army (10).² Each tetrad was composed of four words or phrases descriptive of job performance or personal qualifications from which a supervisor was to select (a) the one most descriptive and (b) the one least descriptive of the individual he was rating. A preliminary investigation had shown that the Army tetrads adapted for use with Public Health Service personnel produced a promising number of scorable alternatives (9).

Section II—Job Proficiency. This was a list of ten major work areas in the Public Health Service from which a supervisor was to indicate the ratee's primary job function. The supervisor was then requested to rate, on a ten-point scale, the quality of the ratee's performance in this function.

Section III—Personal Qualifications. This section consisted of ten-point rating scales for the evaluation of eight personality characteristics such as reaction to criticism, freedom from bias and emo-

² Appreciation is expressed to the Personnel Research Branch, The Adjutant General's Office, Department of the Army, for making these materials available.

tional upset, ability to work with others, ability to act on own responsibility, and diligence and persistence in performing necessary work.

Section IV—Check List. This part consisted of 22 statements which were to be marked as applying or not applying to the ratee. The statements concerned professional knowledge, interest in work, planning and organizing ability, leadership, versatility, and other characteristics related to work performance. In the development of the Check List, some 500 statements had been extracted verbatim from "Remarks" sections of the Officer's Progress Report, placed in 12 logical categories, and sorted according to Thurstone's variation of the method of equal-appearing intervals (8, 11). The 22 statements comprising the final check list were those which showed the least variability (smallest semi-interquartile range) and which were deemed most relevant to performance.

Officer's Progress Report

Samples of parts of the Progress Report used in 1949 to obtain periodic efficiency ratings on commissioned personnel are shown in the Appendix. The Progress Report contains two types of evaluations:

Rating Scales. This section consists of 11 five-point rating scales for evaluating such factors as judgment, general professional knowledge, proficiency in assigned duties, industry, tact, initiative, and dependability. The scales are scored by a point system, odd number values of one through nine being assigned to the five points on each scale; values from all scales are averaged to obtain a total score.

Narrative Comments. Several questions in the Progress Report also elicit narrative comments concerning a ratee's performance. A method for scoring these

comments was developed (8). Its use involves assigning to a comment in a Progress Report the scale value of a matching comment in a scoring manual. The total score for the Narrative Comments is the average of the scale values of all comments in a Report.

Total raw scores from the Narrative Comments and the Rating Scales are separately converted to standard scores on the basis of norms established for each officer grade and profession.³

COLLECTION OF DATA

Criterion ratings were obtained during 1949 from 45 of the 54 Public Health Service installations in the United States, including 14 hospitals, 10 regional offices, 8 divisions, 8 laboratories of the National Institutes of Health, and 5 other installations such as outpatient clinics. Nine stations were excluded from the study because of practical considerations and the small numbers of possible ratees at most of these stations.

Ratings were obtained in a systematic manner by a staff representative who explained the method of rating and administered the forms. Officers who worked together, regardless of profession or grade, met in groups and rated each other. For each of the rating factors, which were randomly alternated, each officer was provided with a roster of all officers at his station. He was then asked to rate each officer, excluding himself and officers he did not know, by placing the ratee at one of the points on the twenty-point scale. Ratings were performed anonymously with the assurance that they were to be used for research purposes only. The lower grades of officers (the equivalents of the Navy ensign through full lieutenant) were rated on

³ Public Health Service officer grades and their Navy equivalents are given below:

Public Health Service	Navy
Junior assistant	Ensign
Assistant	Lieutenant (j.g.)
Senior assistant	Lieutenant
Full	Lieutenant commander
Senior	Commander
Director	Captain

one scale and the higher grades on another in an effort to reduce irrelevant grade-associated factors which might affect the ratings.

One month after the collection of criterion data, copies of the Experimental Efficiency Report, directions for completing, and a schedule for designating supervisors to mark the Experimental Reports were mailed to the officers in charge of the installations which had been visited for purposes of collecting criterion ratings. Two independent Reports were requested on each ratee, one from his immediate officer supervisor and another from either the officer in charge or his representative.

The Officer's Progress Report is requested annually for all officers at the Full grade (Navy lieutenant commander) and above, and semi-annually for all officers below this grade. For purposes of the present study, one Progress Report was selected, where possible, for each officer on whom an Experimental Report had been completed. The Progress Reports selected were those completed within six months before or after the Experimental Reports. In most instances, this time control resulted in a matching of the two Reports on ratee's Service profession, grade, corps, and station. Progress Reports which did not match the Experimental Reports on these factors were eliminated from the study.

DEVELOPMENT OF EXPERIMENTAL REPORT SCORING KEYS

Designation of Occupational Groups for Item Analysis

The diversity of professions and job functions within the Public Health Service necessitated the designation of separate occupational groups for which Experimental Report scoring keys could be developed. An early analysis of the Experimental Report showed that items scorable for various ratee professions did not appreciably overlap (9). In view of these considerations, the following factors were controlled in designating groups for item-analysis purposes.

Criteria. Intercorrelations among scores on the four criteria (not shown in tabular form) revealed that Work Performance and Personality produced the lowest intercorrelations, ranging from .50 to .72; all other correlations were

higher, ranging from .74 to .92. The decision was made to use only the two more independent criteria, Work Performance and Personality, for purposes of item analyzing the Experimental Report.

Station. This factor was controlled by type of station. Stations were classified into three groups according to their major functions: (a) medical care, furnished in installations such as hospitals and outpatient clinics; (b) public health work, carried on in regional offices and in such divisions as those of the Bureau of State Services; and (c) research, performed in installations such as the National Institutes of Health.

Profession of rater. Among medical care personnel (in hospitals and outpatient clinics), where the number of criterion ratings permitted, profession of rater was controlled by using only those ratings performed by members of the ratee's profession. For nurses working in medical care, two groups of raters were used: physicians and nurses. In the public health and the research groups, established by the station control, ratings by all professions of raters were used because of the small numbers of personnel in any one profession. Further, public health and research personnel are usually given efficiency ratings by supervisors working in the same functional area, but not in the same profession. Medical care personnel, however, not only receive efficiency ratings from supervisors in the medical care field, but frequently from those in the particular profession of the ratee.

Profession of ratee. For purposes of item analysis, this variable was controlled among medical care personnel by the establishment of separate groups according to the three major professions represented—medicine, dentistry, and nursing.

TABLE 1
MEANS AND STANDARD DEVIATIONS OF CRITERION SCORES IN EACH OCCUPATIONAL
GROUP DESIGNATED FOR ITEM ANALYSIS

Occupational Group	Work Performance Criterion			Personality Criterion		
	No. Ratees	M ^b	SD	No. Ratees ^c	M	SD
Physicians (1) ^a	158	12.42	2.92	159	11.77	2.88
Physicians (2)	161	12.54	2.72	161	11.84	2.60
Public health personnel (1)	90	13.23	2.43	99	12.75	2.35
Public health personnel (2)	88	13.60	2.34	100	13.23	2.25
Research personnel (1)	66	13.88	2.52	60	12.97	2.33
Research personnel (2)	65	13.79	2.63	67	12.50	3.08
Nurses rated by nurses (1)	56	13.28	2.80	56	12.66	3.04
Nurses rated by nurses (2)	55	13.42	2.64	55	12.84	2.80
Nurses rated by physicians	92	12.44	2.17	91	11.51	2.22
Dentists	60	13.29	1.98	62	12.90	2.10

^a (1) = sample 1; (2) = sample 2.

^b Mean scores from matched samples did not differ significantly at the .05 level or below.

^c In some groups more ratees were evaluated on the Personality than on the Work Performance criterion, perhaps because raters felt they had greater opportunity to observe personal qualifications than job performance.

ing. The groups of public health and research personnel were not broken down by profession of ratee for the same reasons as those specified in the discussion of profession of rater.

Grade of ratee. Grade was controlled by the proportionate representation of each ratee grade in high, middle, and low criterion groups identified within each of the separate item-analysis groups. The criterion groups were the upper 27 per cent, the middle 46 per cent, and the lower 27 per cent of ratees, determined by the average rating they received from the appropriate group of raters on the separate criteria of Work Performance and Personality.

Within the occupational field of medical care, then, four groups were established for purposes of item analysis: (a) physicians rated by physicians; (b) nurses rated by nurses; (c) nurses rated by physicians; and (d) dentists rated by dentists and physicians. Two other item-analysis groups based on occupational fields were also established: (a) public health personnel rated by public health personnel; and (b) research personnel

rated by research personnel.

A criterion score, the average of five or more ratings, was computed for each ratee on each criterion. A minimum of five ratings was required to obtain as highly reliable scores as possible without excluding from the study a large number of the ratees. Where numbers permitted, the groups designated for item analysis were split into matched samples to provide for cross validation. All groups except the dentists and the nurses rated by physicians contained enough ratees to furnish split samples.

The number in each item-analysis group, as well as the mean and the standard deviation of the criterion scores on each rating factor, is shown in Table 1. From Table 1, it is important to note that none of the mean differences in matched groups is significant (five per cent level or below).

Item Analysis

For purposes of item analysis, one Experimental Report was selected for each ratee. The one Report selected, termed the "primary" Report, was gen-

erally completed by a supervisor whose administrative level corresponded to that of a branch chief in a division or a clinical director in a hospital. A "secondary" Report was available on most rates; this Report was generally completed by a supervisor, such as the officer in charge, who was at a higher organizational level than the supervisor completing the primary Report. Secondary Reports, although not used in item analysis, were scored to provide additional validation data.

Section I—Forced Choice. In a previous study, it was found that of four methods for item analyzing forced choice tetrads, the critical-ratio technique appeared the most useful in that it was relatively easy to apply, gave readily interpretable results, and yielded item weights which were in close agreement with weights derived from the other methods studied (9). In the present work, critical ratios were used to item analyze the Forced Choice section against the separate criteria of Work Performance and Personality. Within each item-analysis group, the significance of the difference was tested between the percentage of the high and the percentage of the low criterion groups rated on each alternative. In the medical groups, the samples were sufficiently large that it was possible to use a critical ratio of 1.96 as the standard for scoring. In the remaining smaller occupational groups, an alternative was deemed scorable if the critical ratio were 1.50 or greater. Unitary positive weights, indicating that a significantly higher percentage of a high criterion group than of a low criterion group had been rated on a given alternative, and unitary negative weights, indicating the reverse, were assigned the scorable alternatives in the tetrads. Alternatives which were nondiscriminating received zero weights.

From the scorable tetrads for each item-analysis group, approximately the best 20 were selected to constitute scoring keys. In addition, for each of the item-analysis groups which had been split, a combined sample scoring key was developed. This key was composed of the best 20 tetrads selected from those in which only alternatives having identical scoring weights in the matched samples had been retained. For the group of nurses rated by nurses, only a combined sample key was developed since the matched samples did not individually yield enough scorable items for separate keys.

A total score on the Forced Choice section was obtained by summing all positively weighted alternatives. A previous study indicated that positive weights scoring yielded as valid results as positive plus negative weights scoring on three lengths of keys, one of which was a twenty-tetrad key. The validity of this length of key also compared favorably with that of the other two key lengths studied (5).

Sections II and III—Job Proficiency and Personal Qualifications. Since these two sections consisted of ten-point rating scales, the same methods of item analysis were used for both. First, the discriminatory capacity of the scales was checked by testing the significance of the difference in the mean scale values of high and of low criterion groups. Of the total number of critical ratios computed on the Personal Qualifications scales, considering all item-analysis groups, 70.6 per cent (113 out of 160) were significant at the .05 level or below. On the single Job Proficiency scale, seven of the item analysis groups yielded significant differences (.05 level or below) between upper and lower 27 per cent groups on one or both criteria.

The raw scores of all ratees in each

item-analysis group were used to develop stanine scores in one-half sigma units, with the mean scale value equalling a stanine of five. The stanine scoring resulting for matched samples was so similar that the split samples were recombined to lend greater stability to the stanine scales. Cross validation of the Job Proficiency and Personal Qualifications sections appeared unnecessary in view of the consistency in scoring from one matched sample to another.

The total score on the Personal Qualifications section was the average of the stanines from the eight rating scales. The score on Job Proficiency was the stanine value of the rating given a ratee in his primary job function.

While it would have been desirable to treat separately each of the 10 functions listed in the Job Proficiency section, this was not possible because of the small number of ratees performing each function. All analyses on this section were made without regard to the type of work involved. Specific job functions were not completely masked, however, in view of the method used in establishing item-analysis groups. For example, of the 179 ratees in the public health groups who were rated on the Work Performance criterion, 138 were given ratings on the Experimental Efficiency Report in the primary function described as "operation in a technical or specialized Public Health program."

Section IV—Check List. Critical ratios were computed to test the significance of the difference in the percentages of high and of low criterion groups marked on each alternative. Items which discriminated between the two groups at the .05 level or below were deemed scorable. Considering both criteria and all item-analysis groups, 42.3 per cent (186 out of 440) of the ratios computed reached this level of significance; the scorable

items, however, were not evenly distributed among the various item-analysis groups.

Considering the separate item-analysis groups, the number of scorable items was such that it was feasible to develop scoring keys against the Work Performance criterion in only the medical, public health, and research groups, and against the Personality criterion in only the medical groups. Since these were split groups, the decision was made to include in the scoring keys only those items that reached the required level of significance in both of the matched samples rather than to develop scoring keys for the separate samples.

The total score on this section was the sum of all positively weighted items (those characteristic of a high criterion group).

RESULTS AND INTERPRETATION

Comparison of Forced Choice Scoring Keys

The Forced Choice section of the Experimental Reports from each of the split sample item-analysis groups was scored by three keys: (a) self scoring, developed from item analysis of the sample being scored; (b) cross scoring, developed from item analysis of the matched sample and used for cross validation; and (c) combined sample scoring, based on alternatives that gave the same scoring weights in both samples.

Validity coefficients based on each type of Forced Choice key are presented in Table 2 for the matched sample groups. Data on the three scoring keys are given for secondary Reports as well as for the primary Reports used in item analysis. From the validity coefficients in Table 2, it may first be noted that the coefficients were highly similar from one matched sample to another. In comparisons of matched sample validities, only

TABLE 2
FORCED CHOICE VALIDITY COEFFICIENTS FOR MATCHED SAMPLE GROUPS^a

Occupational Group	Work Performance Criterion				Personality Criterion			
	No. Ratees	Self Scoring	Cross Scoring	Combined Scoring	No. Ratees	Self Scoring	Cross Scoring	Combined Scoring
Primary Reports								
Physicians (1) ^c	158	.67	.65	.66	159	.71 ^d	.61	.69 ^e
Physicians (2)	161	.58	.55	.56	161	.55	.56	.54
Public health personnel (1)	90	.62 ^{df}	.50	.57 ^e	99	.53	.56	.58
Public health personnel (2)	88	.61	.55	.59	100	.55	.56	.56
Research personnel (1)	66	.67	.63	.60 ^e	69	.65	.55	.62
Research personnel (2)	65	.67 ^d	.57	.62	67	.60	.56	.64
Median <i>r</i>		.65	.56	.61		.58	.56	.60
Secondary Reports								
Physicians (1)	120	.67	.66	.67	120	.70	.69	.70
Physicians (2)	123	.70	.67	.68	123	.63	.66	.66
Public health personnel (1)	64	.38 ^f	.34	.30 ^b	71	.47	.49	.52
Public health personnel (2)	63	.47	.44	.45	68	.42	.50	.47
Research personnel (1)	51	.40	.31 ^b	.37	53	.64	.55	.59
Research personnel (2)	56	.53	.52	.57	58	.50	.60	.58
Median <i>r</i>		.50	.48	.51		.57	.58	.59

^a All *rs* not marked are significantly different from zero at the .01 level or below.

^b *r* is significantly different from zero at the .05 level.

^c (1) = sample 1; (2) = sample 2.

^d *r* on self scoring is significantly higher at the .01 level than *r* on cross scoring.

^e *r* on combined scoring is significantly higher at the .05 level than *r* on cross scoring.

^f *r* on self scoring is significantly higher at the .05 level than *r* on combined scoring.

two differences significant at the .05 level or below occurred. These were on primary Reports, Personality criterion, in the comparison of physicians (1) and (2)⁴ on the self (*rs* = .71 vs. .55) and the combined scoring keys (*rs* = .69 vs. .54).⁵

Table 2 also shows that the self and the combined keys, both of which represent the use of scoring keys with item-analysis groups, tended to produce higher validity coefficients than did scoring keys used with independent samples established for cross validation. This trend is apparent both from the median correla-

tions and from the tests of differences in validity from one type of scoring to another.⁶ It should be noted that the differences in validity by type of scoring appeared to decrease when scoring keys were applied to Reports (secondary) independent of those used in item analysis. Only one significant difference in validity by type of scoring occurred on the secondary Reports.

Correlations (not shown in the table) among the three types of scoring keys were high. Considering coefficients based on both primary and secondary Reports,

⁴ Here, (1) = sample 1; (2) = sample 2.

⁵ For purposes of testing the significance of the difference in *rs*, *rs* were transformed to *zs*. Tests of differences reported in this paper involved independent samples and the same sample with one array in common (4, p. 124, formulas 45, 47, and 49).

⁶ Median correlation coefficients have been presented in tables merely to aid the reader in observing trends in the data. They are not intended to be precise summary statistics since an assumption that the various officer groups were samples drawn from a common population is not warranted.

TABLE 3
VALIDITY COEFFICIENTS FOR ALL SECTIONS OF THE EXPERIMENTAL EFFICIENCY REPORT^a

Occupational Group	Work Performance Criterion					Personality Criterion				
	No. Ratees	FC	JP	PQ	CL	No. Ratees	FC	JP	PQ	CL
Primary Reports										
Physicians	310	.61	.62	.58	.54	320	.62	.40	.50	.46
Public health personnel	178	.58	.49	.49	.54	199	.57	.29	.36	
Research personnel	131	.65	.44	.39	.59	136	.63	.18 ^b	.29	
Nurses rated by nurses	111	.42	.31	.39		111	.43	.22 ^b	.35	
Nurses rated by physicians	92	.44	.28	.43		91	.37	.20 ^c	.36	
Dentists	60	.53	.49	.55	.50	62	.35	.41	.40	.33
Median <i>r</i>		.56	.47	.46	.54		.50	.26	.36	.40
Secondary Reports										
Physicians	243	.67	.58	.60	.60	243	.68	.52	.58	.56
Public health personnel	127	.37	.37	.37	.49	139	.50	.30	.33	
Research personnel	107	.48	.47	.41	.49	111	.58	.32	.39	
Nurses rated by nurses	91	.40	.24 ^b	.35		91	.53	.07 ^c	.22 ^b	
Nurses rated by physicians	73	.52	.22 ^c	.34		72	.45	.10 ^c	.27 ^b	
Dentists	42	.66	.55	.48	.64	43	.50	.50	.35 ^b	.49
Median <i>r</i>		.50	.42	.39	.55		.52	.31	.34	.53

^a All *r*s not marked are significantly different from zero at the .01 level or below.

^b *r* is significantly different from zero at the .05 level.

^c *r* does not reach the .05 level of significance.

the median correlation between the self and the cross keys was .92, between the self and the combined, .96, and between the cross and the combined, .96. Both the self and the cross scoring keys correlated highly with the combined key since the latter was composed of the tetrad alternatives scored in both samples.

It was to be expected that the self and the combined keys would yield the higher validity coefficients since they were used to score item-analysis Reports. However, when the self keys were applied as cross keys to Experimental Reports independent of those used in item analysis, the validities exhibited surprisingly little decrease in size. Out of 24 possible comparisons of the self and the

cross keys, only three (12.5 per cent) were significant at the .05 level or below. Although cross-validation data were not available for the combined keys, it seems likely that in successive samples they would have greater stability than keys derived from item analysis of Reports completed on a single sample. For this reason, subsequent discussion of the Forced Choice section of the Experimental Report will be based only on data from the combined sample scoring keys. In order to increase the reliability of the statistics based on these keys, matched samples have been recombined.

Validity of the Experimental Report

Table 3 presents the validity coefficients based on the Forced Choice (FC),

Job Proficiency (JP), Personal Qualifications (PQ), and Check List (CL) sections of the Experimental Report.⁷ To facilitate interpretation of the correlations, the factors which may be influencing variations in the data will be individually considered.

Report sections. From the median correlations in Table 3, it appears that the validity coefficients for the Forced Choice and Check List sections were higher than those obtained for the Job Proficiency and the Personal Qualifications sections. More specific comparisons of those Experimental Report sections which showed significant differences in validity may be seen in Table 4.

Out of the 108 possible comparisons of Experimental Report sections, 36 (33.3 per cent) were significant at the .05 level or below. In 27 of the 36 significantly different pairs of coefficients, the Forced Choice section exhibited higher validities. In only one instance was the validity of the Forced Choice section significantly lower than that of another section. The Job Proficiency and Personal Qualifications scales produced the lowest coefficients; in 15 instances, each of these sections yielded a validity coefficient which was significantly lower than that of another section. The Job Proficiency scale in only three instances and the Personal Qualifications section in only two instances produced validities which were significantly higher than those of other Report sections. The Check List in four comparisons yielded a significantly higher coefficient than another section, and in five comparisons, a significantly lower coefficient. In four of the five instances in which the Check List produced a lower coefficient, it was compared with the Forced Choice section of the Report.

In general, the Forced Choice section gave the highest validity coefficients. Comparisons of validities on the remaining three sections of the Experimental Report yielded relatively few significant differences although the Check List,

where available, tended to produce somewhat higher validities than did the Job Proficiency and Personal Qualifications sections.

TABLE 4

COMPARISONS OF SECTIONS OF THE EXPERIMENTAL REPORT IN WHICH VALIDITY COEFFICIENTS DIFFERED SIGNIFICANTLY

Occupational Group	Primary Reports	Secondary Reports
	Work Performance Criterion	
	Report Sections Compared ^a	Report Sections Compared ^a
Physicians	FC vs. CL ^c JP vs. CL ^c	FC vs. JP ^b FC vs. PQ ^c FC vs. CL ^b
Public health personnel		CL vs. FC ^c
Research personnel	FC vs. JP ^b FC vs. PQ ^b CL vs. JP ^c CL vs. PQ ^b	
Nurses rated by physicians		FC vs. JP ^b FC vs. PQ ^c
Dentists		FC vs. PQ ^c CL vs. PQ ^c
Physicians	Personality Criterion	
	FC vs. JP ^b FC vs. PQ ^b FC vs. CL ^b PQ vs. JP ^b	FC vs. JP ^b FC vs. PQ ^b FC vs. CL ^b JP vs. PQ ^c
Public health personnel	FC vs. JP ^b FC vs. PQ ^b	FC vs. JP ^c FC vs. PQ ^c
Research personnel	FC vs. JP ^b FC vs. PQ ^b PQ vs. JP ^c	FC vs. JP ^c FC vs. PQ ^c
Nurses rated by nurses	FC vs. JP ^c	FC vs. JP ^b FC vs. PQ ^b
Nurses rated by physicians		FC vs. JP ^c
Dentists		JP vs. PQ ^c

⁷ Reports in the dental group were scored by the key developed on physicians. An exploratory validation study showed as high validities for dentists as for physicians when the medical key was used to score reports in both groups.

^a The rating section on which the higher validity was obtained is listed first.

^b Validity coefficients for the sections compared differ significantly at the .01 level or below.

^c Validity coefficients for the sections compared differ significantly at the .05 level.

TABLE 5
COMPARISONS OF OCCUPATIONAL GROUPS IN WHICH VALIDITY COEFFICIENTS
DIFFERED SIGNIFICANTLY

Primary Reports							
Work Performance Criterion			Personality Criterion				
Groups Compared ^a		Report Section	Groups Compared ^a		Report Section		
Physicians	vs.	Nurses rated by nurses	FC ^c	Physicians	vs.	Nurses rated by nurses	FC ^c
		Nurses rated by phys.	FC ^c			Nurses rated by phys.	FC ^b
		P. h. personnel	JP ^c			Dentists	FC ^c
		Res. personnel	JP ^c			Res. personnel	JP ^c
			JP ^b			Res. personnel	PQ ^c
Res. personnel	vs.	Nurses rated by nurses	JP ^b	P. h. personnel	vs.	Nurses rated by phys.	FC ^c
		Nurses rated by phys.	PQ ^c			Dentists	FC ^c
		Res. personnel	PQ ^c			Res. personnel	FC ^c
		Nurses rated by nurses	FC ^c			Nurses rated by nurses	FC ^c
		Nurses rated by phys.	FC ^c			Nurses rated by phys.	FC ^c
				Dentists		Dentists	FC ^c

Secondary Reports							
Physicians	vs.	P. h. personnel	FC ^b	Physicians	vs.	P. h. personnel	FC ^b
		Res. personnel	FC ^c			Nurses rated by phys.	FC ^c
		Nurses rated by nurses	FC ^b			P. h. personnel	JP ^c
		P. h. personnel	JP ^c			Res. personnel	JP ^c
		Nurses rated by nurses	JP ^b			Nurses rated by nurses	JP ^b
		Nurses rated by phys.	JP ^b			Nurses rated by phys.	JP ^b
		Nurses rated by phys.	PQ ^b			P. h. personnel	PQ ^b
		P. h. personnel	PQ ^c			Res. personnel	PQ ^c
		Res. personnel	PQ ^b			Nurses rated by nurses	PQ ^b
		Nurses rated by nurses	PQ ^c			Nurses rated by phys.	PQ ^b
Dentists	vs.	Nurses rated by phys.	FC ^c	Dentists	vs.	Nurses rated by nurses	JP ^c
		P. h. personnel	JP ^c			Nurses rated by phys.	JP ^c

^a The group in which the higher validity was obtained is listed first.
^b Validity coefficients for the groups compared differ significantly at the .01 level or below.
^c Validity coefficients for the groups compared differ significantly at the .05 level.

Occupational group. From Table 3, the significance of the difference was also tested between the validity coefficients obtained from one occupational group to another. The comparisons yielding significant differences are summarized in Table 5. It should be mentioned that tests of differences were only made between independent occupational groups. The two nursing groups, which overlapped in membership, were not compared since the amount of computational work involved did not seem warranted by the relatively small differences in validity that occurred in most instances between the two groups.

Out of the 182 group comparisons made, Table 5 shows that 44 (24.2 per cent) yielded differences significant at the .05 level or below. It is rather striking that in 33 of the significant comparisons, the medical group produced the higher validity coefficient, while in no instance did it produce a significantly lower one. In 25

of the significant comparisons, the lower coefficient occurred in one of the two nursing groups. Coefficients in the public health, research, and dental groups tended to be of about the same magnitude, differing in some instances from the two extreme groups of physicians and nurses. No significant differences were found between the public health and the research groups; only one significant difference occurred in the comparisons of dentists and research personnel, and two in the comparisons of dentists and public health personnel.

The higher validities in the physicians group are perhaps due to factors intrinsic in the work situation. Such factors may be supervisors' relatively greater opportunity to observe carefully the work of medical personnel, particularly interns or lower grade officers under close supervision, and to develop evaluation standards and rating experience since physicians constitute the largest professional group in the Public Health Service.

Another possible explanation of the

validity coefficients obtained for the physicians group is that, in item analysis, a higher critical ratio was used as the standard for scoring in this group than in the other occupational groups. However, a recent study would seem to indicate that a stringent requirement for the level of discrimination of individual items does not necessarily increase total validity (2).

Criteria. Inspection of the validity coefficients (Table 3) from one criterion to another within the same level of supervisor shows that on all but the Forced Choice section higher validities occurred on Work Performance than on Personality. On the Forced Choice section, as high or higher validities were obtained on the Personality as on the Work Performance criterion for both levels of reporting supervisor, and for all occupational groups except the dentists, and the nurses rated by physicians.

A possible explanation of the observed differences in validity by criteria may be that a supervisor, if able to ascertain what constitutes a "good" and a "poor" rating (as is the case for rating scales and check lists), can more objectively evaluate the ratee on observable work-performance characteristics than he can on personal characteristics. When good and poor ratings are not so readily discernible, as presumably is the case with the forced choice type of evaluation, the objectivity of evaluations is perhaps increased so that they are as valid measures of the factors involved in a Personality as in a Work Performance criterion.

Level of supervisor. The validity coefficients in Table 3 may also be compared by level of supervisor. Since the primary Reports were used in item analysis, it is to be expected that they would yield higher validities than the secondary Reports. The median correlations in the

table, however, indicate that validity held up surprisingly well on secondary Reports; the median correlations on the Personality criterion were even somewhat higher on the secondary than on the primary Reports. Considering the 42 pairs of coefficients which can be compared from one level of supervisor to another, primary Reports produced the higher validity in 21 comparisons, and secondary Reports produced the higher validity in the same number of comparisons. The median difference in coefficients was .07 for those comparisons in which primary Reports produced higher correlations, and .08 for those comparisons in which secondary Reports gave higher validities.

Differences in validity by level of supervisor, however, were observable within specific occupational groups. In 75 per cent or more of the comparisons of coefficients within the public health and the nurses-rated-by-nurses groups, the higher validity occurred on primary Reports. Since the primary Reports were used in item analysis, this finding is in the expected direction but in addition possibly reflects the fact that the primary, more immediate, supervisors of these groups are more likely to be in the ratees' profession than are the secondary supervisors. In 75 per cent or more of the comparisons within the medical and dental groups, the higher coefficients occurred on secondary Reports. It should be mentioned that hospitals and outpatient clinics are administered by a medical officer, and dental services are headed by a dental officer. For this reason, both the secondary and the primary supervisors of physicians and dentists are likely to be in either the same profession as the ratees or the same as the raters who performed criterion ratings. Further, secondary super-

visors of physicians and dentists are likely to be officers who routinely review efficiency reports and, therefore, have more information available as a basis for evaluating a ratee than do the primary supervisors.

In general, then, Reports completed independently by a second group of supervisors produced validities that compared favorably with those based on the Reports used in item analysis. It is likely that, had the secondary supervisors been at the same administrative level as the primary, differences in validity by level of supervisor which were apparent in specific occupational groups would have tended to occur less often; that is, the validities would be more nearly the same in all groups than occurred in the present data.

Ratee grade. A control on ratee grade was used both in the administration of criterion rating forms and in the establishment of occupational groups for purposes of item analysis. Nonetheless, since it was not feasible to control grade more precisely, this factor may be operating to increase spuriously the correlations shown in Table 3. In order to check this possibility, validity coefficients by grade were computed; small numbers of ratees made it necessary in some instances, however, to combine adjacent grades such as the senior and the director. Correlations by grade are presented in Table 6.

If grade were a systematic factor affecting both the criterion and Experimental Report variables, it would be expected that the validities based on all grades would be higher than the individual grade validities. That this is not the case may be seen from Table 6. There does not appear to be any consistent trend in the correlations as a function of grade level. The effect of combining grades was

the masking of the higher validity obtained for some specific grades. In only one instance, in the nurses-rated-by-physicians group, was the correlation for all grades higher than any of the validity coefficients for the individual grades.

The significance of the difference in validity coefficients from one grade to another was tested. The relatively few comparisons, 16 out of a possible 122 (13.1 per cent), which yielded differences significant at the .05 level or below are shown in Table 6. No differences occurred in the public health group, and only two were found in the group of nurses rated by physicians.

In the medical group, the significant differences tended to involve higher validities in the Assistant grade as compared with other grades. Out of the 14 significant differences found in this professional group, 10 occurred in comparisons in which the higher coefficient was in the Assistant grade. This finding may be due to the fact that the majority of Assistant grade physicians are interns under close supervision; the supervisors of interns are experienced raters who have ample opportunity to observe the interns' performance.

The remaining four significant differences in the medical group occurred on the Reports completed by the secondary supervisors, Work Performance criterion, in the comparison of the combined Senior and Director grade with the Senior Assistant grade. The unusually high validities in the combined Senior and Director grade may have been due to the small number of ratees (about half the number available on primary Reports) and perhaps to a selective factor that resulted in the designation of highly experienced raters as the secondary supervisors for this grade.

Although differences in the validity coefficients found for the various ratee grades did occur, they were presumably the result of certain identifiable influences. The grade factor, as such, does not appear to have been operating in any systematic manner which spuriously increased the validities based on all grades.

Reliability of the Experimental Report
Rater agreement. Correlations between scores from primary and secondary Reports, shown in Table 7, provide one

TABLE 6
VALIDITY COEFFICIENTS FOR THE EXPERIMENTAL EFFICIENCY REPORT
BASED ON SEPARATE GRADES

Occupational Group	Grade	Primary Reports									
		Work Performance Criterion					Personality Criterion				
		N	FC	JP	PQ	CL	N	FC	JP	PQ	CL
Nurses rated by physicians	Assistant	48	.64 ^a	.45	.04		48	.39	.20	.32	
	Senior assistant	28	<u>.23</u>	.02	.19		28	.39	.20	.52	
	All grades ^a	92	.44	.28	.43		91	.37	.20	.36	
Physicians	Assistant	83	.64	.71 ^c	.65	.66 ^c	83	.62	.55 ^c	.59	.57 ^b
	Senior assistant	108	.54	<u>.53</u>	.50	<u>.45</u>	108	.65	.35	.42	.46
	Full	58	.59	<u>.45</u>	.47	.45	58	.65	.34	.54	.45
	Senior & Director	70	.61	.65	.57	.48	71	.50	<u>.24</u>	.42	<u>.20</u>
	All grades	319	.61	.62	.58	.54	320	.62	.40	.50	.46
Public health personnel	Full	55	.68	.63	.55	.66	59	.59	.39	.30	
	Senior	63	.61	.43	.45	.52	73	.61	.35	.38	
	All grades	178	.58	.49	.49	.54	199	.57	.29	.36	
Secondary Reports											
Nurses rated by physicians	Assistant	35	.63	.13	.39		35	.51	<u>-.20</u>	.18	
	Senior assistant	24	.53	.49	.22		24	.48	<u>.58^b</u>	.50	
	All grades	73	.52	.22	.34		72	.45	.10	.27	
Physicians	Assistant	62	.69	.70 ^c	.68 ^c	.70 ^c	62	.66	.69 ^c	.68	.64
	Senior assistant	100	<u>.58</u>	<u>.44</u>	<u>.42</u>	<u>.44</u>	100	.67	<u>.44</u>	.54	.50
	Full	50	.73	.52	.62	.58	50	.72	<u>.43</u>	.49	.48
	Senior & Director	31	.80 ^c	.73 ^c	.75 ^c	.80 ^b	31	.68	.51	.62	.74
	All grades	243	.67	.58	.60	.60	243	.68	.52	.58	.56
Public health personnel	Full	46	.46	.56	.50	.51	49	.53	.48	.56	
	Senior	42	.44	.34	.44	.55	47	.65	.24	.39	
	All grades	127	.37	.37	.37	.49	139	.50	.30	.33	

^a The N for All Grades is based on all available cases including those in grades too small for the separate grade analysis.
^b r is significantly higher at the .01 level than the underlined r in same column within the same report within the same officer group.

^c r is significantly higher at the .05 level than the underlined r in same column within the same report within the same officer group.

TABLE 7
CORRELATIONS BETWEEN SCORES FROM PRIMARY AND SECONDARY REPORTS^a

Occupational Group	Work Performance Criterion					Personality Criterion				
	No. Ratees	FC	JP	PQ	CL	No. Ratees	FC	JP	PQ	CL
Physicians	243	.57	.51	.58	.61	243	.60	.51	.58	.59
Public health personnel	127	.57	.37	.42	.47	139	.63	.46	.41	
Research personnel	107	.59	.53	.62	.54	111	.64	.54	.61	
Nurses rated by nurses	91	.59	.39	.55		91	.57	.39	.55	
Nurses rated by physicians	73	.50	.27 ^b	.48		72	.62	.31	.48	.61
Dentists	42	.65	.55	.52	.58	43	.64	.55	.52	
Median r		.58	.45	.54	.56		.63	.49	.54	.60

^a All rs not marked are significantly different from zero at the .01 level or below.

^b r is significantly different from zero at the .05 level.

TABLE 8
RELIABILITY COEFFICIENTS FOR THREE SECTIONS OF THE EXPERIMENTAL
EFFICIENCY REPORT

Occupational Group	Forced Choice					Personal Qualifications ^a		Check List		
	No. Rates	No. Scored Tetrads	No. Scored Alternatives ^b	$r_{\frac{1}{2}I/II}$	r_{II}	$r_{\frac{1}{2}I/II}$	r_{II}	No. Scored Alternatives	$r_{\frac{1}{2}I/II}$	r_{II}
Physicians	310	20	28	.83	.91	.80	.94	12	.73	.85
P. h. personnel	178	18	30	.83	.91	.83	.91	7	.67	.80
Res. personnel	131	18	22	.78	.88	.90	.95	9	.68	.81
Nurses rated by nurses	111	21	20	.78	.88	.92	.96			
Nurses rated by phys.	92	19	23	.64	.78	.94	.97	12	.86	.92
Dentists	60	20	28	.83	.91	.90	.95			
Median r_{II}					.90		.95			.83

^a The number of rating scales was eight, with the possible total stanine score ranging from 8 to 72.

^b The number of scored alternatives rather than tetrads was used as the basis for computing reliability coefficients.

kind of measure of the reliability of the Experimental Report.

Correlations between Reports completed by the two groups of supervisors ranged from .27 to .65; over half were .55 or higher. Only the lowest correlation, .27, failed to reach the .01 level of significance; it was significant at the .05 level. It should be noted that the Job Proficiency (JP) section which consisted of a single rating scale tended to produce the lowest correlations. As the median r s indicate, the Forced Choice (FC) section tended to yield the highest correlations, although for the Work Performance criterion these did not differ markedly from those produced by the Personal Qualifications (PQ) and Check List (CL) sections. For the Personality criterion, the Forced Choice section gave the highest correlations in all instances, although the two correlations available on the Check List were of comparable size.

Since the primary and the secondary supervisors were at different administrative levels, the correlation coefficients are lower than might be obtained between Reports completed by different supervisors at the same level or between Reports completed by the same supervisor on two different occasions. Considering the factors operating to lower the coeffi-

cients, the correlations between scores on primary and secondary Reports, as measures of rater agreement, are fairly high.

Spearman-Brown estimates. From primary Reports, scored by the key developed against the Work Performance criterion, correlations between the odd and the even alternatives in each of three rating sections were corrected for length by the Spearman-Brown formula. A previous paper has reported that for Forced Choice tetrads, Spearman-Brown estimates of reliability were fairly close approximations of empirical reliabilities (5). Spearman-Brown estimates based on the present data are shown in Table 8. Since the Job Proficiency section involved only a single rating scale, it was not possible to compute a split-half coefficient on this part of the Report.

Considering the length of the various sections of the Report, the reliability coefficients are in the high range. As can be seen from Table 8, the number of scored alternatives on the Forced Choice and Check List sections varied somewhat for the different occupational groups. The number of rating scales in the Personal Qualifications section was the same (eight) for all groups.

The highest Spearman-Brown esti-

TABLE 9

COMPARISON OF EMPIRICAL VALIDITY COEFFICIENTS WITH VALIDITIES PREDICTED ON THE BASIS OF AN INCREASE IN LENGTH OF SCORING KEY

Occupational Group	Section	No. Scored Alternatives	r_{11}	Empirical Validity	Estimated Validity ^c	Limit of Validity ^d
Physicians	FC	28	.91	.61 ^b		.64
	PQ	8	.94	.58	.59	.60
	CL	12	.85	<u>.54</u>	.56	.59
Public health personnel	FC	30	.91	.58		.61
	PQ	8	.91	.49	.51	.51
	CL	7	.80	.54	.59	.60
Research personnel	FC	22	.88	.65 ^a		.69
	PQ	8	.95	<u>.39</u>	.40	.40
	CL	9	.81	.59 ^a	.63	.66
Nurses rated by nurses	FC	29	.88	.42		.45
	PQ	8	.96	.39	.40	.40
Nurses rated by physicians	FC	23	.78	.44		.50
	PQ	8	.97	.43	.43	.44
Dentists	FC	28	.91	.53		.56
	PQ	8	.95	.55	.56	.56
	CL	12	.92	.50	.51	.52

^a r is significantly higher at the .01 level than the underlined r in the same column within the same officer group.

^b r is significantly higher at the .05 level than the underlined r in the same column within the same officer group.

^c Estimated validity based on the same number of scored alternatives as the Forced Choice section.

^d Limit of validity if report section were made infinitely long.

mates of reliability occurred on the Personal Qualifications (PQ) scales; all r_{11} 's were .91 or higher. Coefficients on the Check List (CL) ranged from .80 to .92, with a median of .83. The Forced Choice (FC) section yielded satisfactory reliabilities in all officer groups except the nurses-rated-by-physicians ($r_{11} = .78$); all other coefficients on this section were .88 or .91, with a median of .90.

Validity as Related to Length of Scoring Key

Since the sections of the Experimental

Report were not equated in length, the fact that the Forced Choice was the longest section may account for its higher validity. The effect of length of scoring key on validity was tested on the Experimental Report sections for which both validity and reliability data were available from the primary Reports scored by keys developed on the basis of the Work Performance criterion. The results of the tests are shown in Table 9 which presents: (a) empirical validities and reliabilities and the number of scored alternatives in each Report section, re-

peated from previous tables for ease of comparison; (b) estimated validities for the Personal Qualifications (PQ) and the Check List (CL) sections, based on an increase in length of these sections to that of the Forced Choice (FC); and (c) the maximum validity that theoretically could be obtained on each section if it were made infinitely long (1, p. 166).

From Table 9, it may be seen that the Forced Choice section produced the highest empirical validity in all but one occupational group, the dental. Theoretically increasing the length of the other Report sections to that of the Forced Choice resulted in one additional group, public health personnel, in which the validity (estimated) of a section other than the Forced Choice was the highest. It is likely, however, that had the Report sections been equated for length, the three significant differences in validity (Table 9, footnotes a and b) that were obtained from one Report section to another would still have occurred.

If each of the Report sections were made infinitely long, it is apparent from a comparison of obtained validities with the theoretical limits of validity in Table 9 that the Personal Qualifications (PQ) section could be expected to show the smallest increase in validity (no more than .02). On the Forced Choice (FC) section, however, the increase to be expected ranges from .03 to .06, and on the Check List (CL), from .02 to .07.

While length of the various Report sections appears to have affected the magnitude of the validity coefficients, the greater number of scored alternatives in the Forced Choice (FC) section does not appear primarily responsible for the generally higher validity of this section. Although the Check List (CL) was the section most affected by the small number of scored alternatives, the evidence on limit of validity seems to indicate the relatively higher validity of the Forced Choice section.

Multiple Correlations

In order to determine the combination of Experimental Efficiency Report sections which would, for each occupational group, maximally predict the separate criteria, multiple correlation coefficients based on primary Reports were computed by the Wherry-Doolittle method of test selection (4, Chap. XIV). The intercorrelations on which the multiple correlation work was based are shown in Table 10. As can be seen from Table 10, the intercorrelations among Report sections were quite high, particularly in the medical and dental groups. In all occupational groups, the highest intercorrelations tended to occur between the Job Proficiency (JP) and Personal Qualifications (PQ) sections, with correlations ranging from .66 to .85. The Forced Choice (FC) and the Check List (CL) sections were also highly related, with a range in correlations from .68 to .82.

The multiple correlational data are presented in Table 11, which shows the \bar{R} s based on selected predictors, the \bar{R} s obtained by application of the Wherry shrinkage formula, the validity coefficient of the best predictor in each occupational group, and the beta weights of predictors in the order in which the predictors were selected.

As Table 11 indicates, all \bar{R} s were significantly different from zero at the .01 level or below. Within each occupational group, the increase in validity as a result of using a team of Report sections rather than a single predictor may be seen by comparing the \bar{R} with the r of the first selected predictor. However, since the Wherry-Doolittle method of test selection does not guarantee that the increment due to the selection of successive variables significantly increases validity, the null hypothesis was tested by use of the F ratio (3, p. 55). Tests for the significance

TABLE 10
INTERCORRELATIONS AMONG SECTIONS OF THE PRIMARY REPORTS^a

Occupational Group	Work Performance Criterion							Personality Criterion						
	N	FC-JP	FC-PQ	FC-CL	JP-PQ	JP-CL	PQ-CL	N	FC-JP	FC-PQ	FC-CL	JP-PQ	JP-CL	PQ-CL
Physicians	319	.64	.70	.79	.85	.69	.73	320	.53	.68	.74	.85	.65	.73
P. h. personnel	178	.56	.56	.68	.78	.59	.49	199	.52	.52		.77		
Res. personnel	131	.51	.47	.77	.79	.48	.45	136	.38	.45		.80		
Nurses rated by nurses	111	.41			.68			111	.34	.56		.68		
Nurses rated by phys.	92	.48	.66		.66			91	.33	.58		.66		
Dentists	60	.73	.70	.82	.80	.75	.76	62	.71	.63	.75	.80	.76	.74

^a All *r*s are significantly different from zero at the .01 level or below.

of the difference in the R^2 (or r^2) based on the first selected predictor and the R^2 based on all selected predictors showed that, on the Work Performance criterion, validity was significantly increased at the .05 level or below in all groups except the dental. On the Personality criterion, a significant increase at the .05 level or below occurred in only the largest group, the medical.

As was previously noted in the discussion of the validity coefficients in Table 3, the data in Table 11 indicate that the validities obtained against the Work Performance criterion were higher than those on Personality, and that the physicians group was the one in which prediction was best.

The relative effectiveness of the four Report sections as measures of performance within each occupational group is evident from the results of the test selection. The predictors in Table 11 are those which, as determined by the Wherry-Doolittle method, combine to produce maximum multiple correlations. Inspection of the order in which predictors were selected and of the number of occupational groups in which they were selected shows that, considering both criteria, the Forced Choice (FC) section was a selected predictor in eleven of the twelve groups. Further, the Forced Choice section was the first selected predictor in nine of the groups. The Personal Qualifications (PQ) section was the next most frequently selected Report section occurring in eight of the officer groups, while the Job Proficiency (JP) section and the Check List (CL) each occurred in three groups. In general, the Forced Choice section in combination with one of the rating scale sections, usually Personal Qualifications, tended to produce the maximum correlations with the criteria.

TABLE II
MULTIPLE CORRELATIONS BASED ON PRIMARY REPORTS

Occupational Group	Work Performance Criterion					Personality Criterion				
	No. Ratees	Selected Pre-dictors	Beta Weights	r of First Pre-dictor	R	\bar{R}^a	No. Ratees	Selected Pre-dictors	Beta Weights	r of First Pre-dictor
Physicians	319	JP FC	.39 .36	.621	.680 ^c	.679	320	FC PQ CL	.57 .10 -.10	.624
Public health personnel	178	FC PQ CL	.31 .21 .23	.580	.637 ^c	.632	199	FC PQ	.52 .09	.571
Research personnel	131	FC CL JP	.43 .20 .13	.650	.674 ^b	.668	136	FC	.63	.631
Nurses rated by nurses	111	FC PQ	.29 .22	.420	.455 ^b	.447	111	FC PQ	.34 .16	.430
Nurses rated by phys.	92	FC PQ	.28 .25	.442	.482 ^b	.472	91	FC PQ	.23 .23	.367
Dentists	60	PQ FC	.34 .30	.545	.584	.575	62	JP	.41	.413

^a All \bar{R} s are significantly different from zero at the .01 level or below.

^b R is significantly higher than r at the .05 level.

^c R is significantly higher than r at the .01 level or below.

Since the multiple correlational work was based on validity coefficients obtained for item-analysis samples, the findings concerning the relative effectiveness of the various Report sections and the sizes of the multiple correlation coefficients require verification on independent samples. Evidence from the "cross" scoring keys, however, indicated that little decrease is to be expected in cross validation of the Forced Choice section of the Report. Further, cross validation of the Personal Qualifications and Job Proficiency scales was not deemed necessary since matched samples had produced highly similar scoring keys. In view of these considerations, it is likely that the validity coefficients will not show a marked decrease in subsequent samples.

Validity of the Officer's Progress Report

Validity coefficients based on the Rating Scales (RS) and the Narrative Comments (NC) sections of the Officer's Progress Report and on sections of the Experimental Report completed by the primary supervisors are shown in Table 12. Considerable attrition in the number of Experimental Reports occurred as a result of using only those rates on whom both Reports were available. The median correlations for the Experimental Report shown in Table 12, however, are about the same size as the corresponding median validities in Table 3.

Tests of the significance of the difference in the validity coefficients in Table 12 from one Report section to another revealed that coefficients in 20 per cent (30 out of a possible 150) of the comparisons differed significantly at the .05 level or below. The specific comparisons which produced significant differences are shown in Table 13. The percentage of significant comparisons was less than occurred in the tests of differences on the Experimental Report (see Table 4); this was probably due to the smaller number of cases available on the combined Reports.

As was previously found, the Forced Choice

TABLE 12
VALIDITY COEFFICIENTS FOR THE EXPERIMENTAL REPORT AND THE OFFICER'S PROGRESS REPORT^b

Occupational Group	Work Performance Criterion							Personality Criterion						
	N ^a	Experimental Report (Primary Reports)					Progress Report	N ^a	Experimental Report (Primary Reports)					Progress Report
		FC	JP	PQ	CL	RS			NC	FC	JP	PQ	CL	
Physicians	187	.62	.69	.61	.54	.55	.57	187	.59	.52	.56	.47	.52	.53
P. h. personnel	108	.64	.49	.47	.50	.56	.52	120	.59	.23 ^c	.30		.32	.33
Res. personnel	74	.74	.56	.49	.65	.58	.69	73	.72	.15 ^d	.30		.35	.33
Nurses rated by nurses	88	.43	.34	.50		.40	.20 ^d	88	.46	.24 ^e	.38		.34	.17 ^d
Nurses rated by phys.	72	.40	.34	.52		.57	.26 ^e	71	.38	.19 ^d	.39		.27 ^e	.15 ^d
Dentists	40	.52	.47	.57	.50	.48	.46	42	.42	.47	.44	.36 ^e	.41	.44
Median <i>r</i>		.57	.48	.51	.55	.56	.49		.53	.24	.39	.42	.35	.33

^a Based on the number of officers on whom both Experimental and Progress Reports were available.

^b All r s not marked are significantly different from zero at the .01 level or below.

^c r is significantly different from zero at the .05 level. ^d r does not reach the .05 level of significance.

TABLE 13

COMPARISONS OF SECTIONS OF THE EXPERIMENTAL REPORT AND THE OFFICER'S PROGRESS REPORT IN WHICH VALIDITY COEFFICIENTS DIFFERED SIGNIFICANTLY

Occupational Group	Work Performance Criterion	Personality Criterion
	Report Sections Compared ^a	Report Sections Compared ^a
Physicians	FC vs. CL ^c JP vs. PQ ^b JP vs. CL ^b JP vs. RS ^b JP vs. NC ^c	FC vs. CL ^b PQ vs. CL ^c
Public health personnel	FC vs. JP ^c FC vs. PQ ^c	FC vs. JP ^b FC vs. PQ ^b FC vs. RS ^b FC vs. NC ^b
Research personnel	FC vs. JP ^c FC vs. PQ ^b FC vs. RS ^c NC vs. PQ ^c	FC vs. JP ^b FC vs. PQ ^b FC vs. RS ^b FC vs. NC ^b PQ vs. JP ^c
Nurses rated by nurses	FC vs. NC ^c PQ vs. JP ^c PQ vs. NC ^b	FC vs. NC ^c
Nurses rated by physicians	FC vs. NC ^c RS vs. JP ^c PQ vs. NC ^c RS vs. NC ^b	

^a The Report section on which the higher validity occurred is listed first.

^b Validity coefficients for the sections compared differ significantly at the .01 level or below.

^c Validity coefficients for the sections compared differ significantly at the .05 level.

(FC) section produced more (18 out of 30) of the significantly higher validity coefficients than any other Report section. In no instance was a validity coefficient on the Forced Choice section significantly lower than that of another section. The number of comparisons in which each of the remaining five Report sections produced a validity significantly higher than another section ranged from none on the Check List to five on the Personal Qualifications section.

From the results of the significance of difference tests and from the median validity coefficients, it is interesting to note that the two sections of the Progress Report, Rating Scales (RS) and Narra-

tive Comments (NC), produced validities that compare favorably with all sections of the Experimental Report except the Forced Choice.

The data on the Officer's Progress Report again suggest the relative superiority of the forced choice type of evaluation as compared with more conventional rating methods. However, since the Progress Report was completed under operational rather than experimental conditions, no attempt will be made to compare the two Reports by use of multiple correlational techniques. It is anticipated that in a later study, it will be possible to collect data on the Progress Report along with cross-validation data for the Experimental Report so that a more intensive comparison of the two Reports can be made.

SUMMARY

This study has compared the relative efficacy of the forced choice technique with other more conventional evaluation methods as measures of the performance of professional health personnel working as commissioned officers in the United States Public Health Service.

Four sections of an Experimental Efficiency Report were studied: (a) 50 Forced Choice tetrads adapted from items developed by the Department of the Army; (b) a ten-point scale for rating a ratee's Job Proficiency in his primary job function; (c) eight ten-point scales for the evaluation of Personal Qualifications; and (d) a twenty-two-item Check List developed from comments appearing in the Officer's Progress Report, the efficiency report in operational use in the Service. In addition, two sections from the Officer's Progress Report were available for comparison with those in the Experimental Report: (a) eleven five-point Rating Scales for evaluating

various aspects of performance in the Public Health Service; and (b) Narrative Comments coded and scored by a method previously developed.

The criteria of Service performance were twenty-point graphic rating scales for the evaluation of Work Performance and Personality. A ratee's criterion score was the average of the ratings given him by his work associates on each criterion. The results of the study have shown that:

1. The Forced Choice section of the Experimental Report was highly effective for evaluating the performance of professional personnel commissioned in the Public Health Service.

Of 24 validity coefficients based on scoring keys developed by selecting the best tetrads from those which had the same empirically determined scoring weights in independent matched samples, 41.7 per cent were .62 or higher. All except one of the coefficients were significant at the .01 level or below; this one was significant at the .05 level (Table 2, "combined scoring").

Only 12.5 per cent of 24 validity coefficients based on item-analysis samples showed a significant decrease at the .05 level or below in cross validation (Table 2, comparison of "self" and "cross" scoring).

2. The validity of the Forced Choice section was generally higher than that of the other Report sections studied.

Out of 36 significant differences (.05 level or below) obtained in comparisons of the validity of the Experimental Report sections, 27 (75 per cent) involved higher validities on the Forced Choice tetrads, while only one involved a lower coefficient on this section (Table 4).

The Forced Choice section contained a greater number of scored alternatives than the other sections of the Experimental Report; estimates of validity based on theoretically making each section infinitely long, however, seemed to indicate that the length of the Forced Choice section was not primarily responsible for its generally higher validity (Table 9).

Out of 12 multiple correlation coefficients computed on the Experimental Report by the Wherry-Doolittle method of test selection, 11 involved the Forced Choice section as a selected predictor; in nine of the 11, this section was the first selected predictor (Table 11).

Comparisons of the validities of sections of

both the Experimental Report and the Officer's Progress Report revealed 30 significant differences; 18 (60 per cent) involved higher validities on the Forced Choice tetrads while none involved a lower coefficient on this section (Table 13).

3. Of six occupational groups for which separate scoring keys were developed for the Experimental Report, the largest group, that of hospital physicians, was the one in which the highest validity coefficients generally occurred. The occupational groups, other than physicians, which were involved in the study were dentists, public health personnel, research personnel, and nurses rated by two different criterion rater groups, physicians and nurses.

Of 44 significant differences (.05 level or below) obtained in comparisons of validity coefficients from one occupational group to another on sections of the Experimental Report, 33 (75 per cent) involved higher coefficients in the medical group (Table 5).

Multiple correlations (\bar{R}) for the Experimental Report computed by the Wherry-Doolittle method were, for the medical group, .68 and .63 against the Work Performance and the Personality criteria, respectively. Both coefficients were significant at the .01 level or below, and both represented a significant increase (.05 level or below) in validity over that obtained on the second-best single Report section. Multiple correlations in the public health and the research groups were also relatively high, ranging from .57 to .67 on the two criteria (Table 11).

4. Validity coefficients were generally higher when Work Performance rather than Personality was used as the criterion.

On all sections of the Experimental Report except the Forced Choice, higher validities were obtained with the Work Performance criterion than with the Personality criterion. Forced Choice validities were not consistently higher for either criterion (Table 3).

Within each officer group, a higher multiple correlation coefficient was obtained for the Experimental Report when Work Performance was used as the criterion than when Personality was used (Table 11).

Considering sections from both the Experimental and the Progress Reports, higher

validities occurred in all but one instance when Work Performance rather than Personality was used as the criterion (Table 12).

5. Experimental Reports completed by a group of supervisors independent of and at a higher administrative level than those completing the Reports used in item analysis produced validities that compared favorably with the validities of the item-analysis Reports.

Of 42 comparisons of validity from one level of supervisor to another, 21 involved higher validity coefficients on item-analysis Reports, and 21 involved higher validities on Reports completed by an independent group of supervisors. The median difference in validity coefficients in those comparisons in which item-analysis Reports yielded the higher validities was .07, and in those in which the independent Reports gave higher coefficients, .08 (Table 3).

6. Validity coefficients based on sections of the Experimental Report did not show a consistent trend as a function of grade level.

Of 122 possible comparisons of validity coefficients from one grade to another, only 16 (13.1 per cent) yielded differences significant at the .05 level or below. Validity coefficients for separate grades were also compared with those based on all grades. The effect of combining grades appeared to be the masking of the higher validity obtained in certain specific grades; in only one instance was a combined grade validity higher than any of the coefficients for the separate grades (Table 6).

7. The sections of the Experimental Efficiency Report exhibited satisfactory reliabilities.

Spearman-Brown estimates of reliability for three of the Report sections ranged from .78 to .97. Median reliabilities (r_{11}) were .95, .90, and .83, respectively, for the Personal Qualifications, the Forced Choice, and the Check List sections. It was not possible to compute a split-half coefficient for the Job Proficiency section since it consisted of a single rating scale (Table 8).

As a measure of rater agreement, scores on Reports completed by two groups of supervisors at different administrative levels were correlated. Over half of the correlations between the two sets of Reports were .55 or higher (Table 7).

8. The Rating Scales and Narrative

Comments sections of the Officer's Progress Report appeared to be about as adequate measures of performance as sections of the Experimental Report other than the Forced Choice.

Median validity coefficients for the Progress Report compared favorably with those for sections of the Experimental Report other than the Forced Choice. Data on the two Reports, however, were collected under different conditions so that comparative results are viewed as tentative (Table 12).

The significance of the difference was tested in the validity coefficients obtained for the various sections of both Reports. Significantly higher (.05 level or below) validities occurred on each section of the Progress Report about as frequently as on the Experimental Report sections other than the Forced Choice (Table 13).

9. Multiple correlations computed on the Experimental Report indicated that prediction was in some instances, but not in others, increased by the use of more than one Report section.

All multiple correlations were significantly different from zero at the .01 level or below. Five of the six correlations based on the Work Performance criterion represented a significant increase (.05 level or below) in validity over that obtained on the best single Report section for each officer group. Only one of those based on the Personality criterion, however, showed such a significant increase (Table 11).

10. The combination of sections of the Experimental Report which produced the maximum correlation with the criteria, as determined by the Wherry-Doolittle method, differed for each of the officer groups studied, but tended to include the Forced Choice in combination with one of the rating scale sections, usually Personal Qualifications.

Of 12 multiple correlations computed, six involved the Forced Choice and Personal Qualifications sections as the only selected predictors, and three involved these two sections in combination with a third section. In one multiple correlation the Forced Choice section was selected in combination with the Job Proficiency scale, and each of these sections was the only predictor selected in the two remaining multiple correlations (Table 11).

IMPLICATIONS OF THE FINDINGS

Evaluation of the performance of highly trained professional personnel poses a difficult measurement problem. The complexities of the work requirements for such personnel make adequate, objective criteria of professional competency difficult to obtain at the present time. Any criterion or criteria should presumably reflect such personal characteristics as professional knowledge, judgment, technical skill, originality, emotional adjustment, and ability to administer programs in a professional specialty. While the inadequacies of the type of criterion employed in the present work are recognized, practical considerations necessitated the use of a conventional work-associates' rating method.

With the type of item analysis and control of experimental variables used in this study, it would appear that, within the limitations imposed by a rating criterion, satisfactory validity and reliability of performance evaluation methods for professional health personnel can be obtained. Of particular interest are the results obtained for the forced choice tetrads which, under the conditions of this study, generally produced

higher validity coefficients than other methods of assessing or reporting efficiency. Since rating-scale methods of efficiency reporting have widespread usage, it may also be of general interest that these methods produced satisfactory validity as measures of professional performance.

The findings appear to be applicable to other organizations employing medical, scientific, and other health personnel similar to those employed by the Public Health Service. With regard to the forced choice items, it may be recalled that the items used here, although scored by keys developed from item analysis of Experimental Efficiency Reports completed on Public Health Service personnel, were developed in another organization on an employee population quite different from that of the Public Health Service. From the evidence concerning validity of the tetrads in the variety of work activities in the Public Health Service (medical care, research, and public health), it may be inferred that the item content and the technique are such as to be relevant in a number of different employment situations.

A. SAMPLES OF ITEMS FROM SECTIONS OF THE EXPERIMENTAL EFFICIENCY REPORT

Directions for completing: From each of the following sets of four words or phrases, mark the one word or phrase in each set which is "most descriptive" and the one which is "least descriptive" of the officer you are rating.

	Most	Least
A. Cannot assume responsibility		
B. Knows how and when to delegate authority		
C. Offers suggestions		
D. Too easily changes his ideas		

	Most	Least
A. Modest and reserved		
B. Doesn't have the drive or force he should		
C. Antisocial		
D. Respected by all fellow officers		

Section II. Job Proficiency

Directions for completing: From the Service functions listed below, select the one you consider to be the primary job of the officer you are evaluating. Rate the officer's job proficiency in this function by marking a position on the ten-point scale.

Directions for completing: From the Service functions listed below, select the one you consider to be the primary job of the officer you

are evaluating. Rate the officer's job proficiency in this function by marking a position on the ten-point scale.

- | | 1 | 2 | 3 |
|--|---|---|---|
| 1. Operation in a technical or specialized Public Health program | | | |
| 2. Care of patients or furnishing services to patients | | | |
| 3. Administration of a clinical or medical care program at any level | | | |
| 4. Directly performing research work | | | |

FOR RATING OFFICER

☐ Number of Function

1 2 3 4 5 6 7 8 9 10

Section III. Personal Qualifications

Directions for completing: By marking a position on a ten-point scale, rate the officer on each of the following personal qualifications.

	1	2	3	4	5	6	7	8	9	10
--	---	---	---	---	---	---	---	---	---	----

Directions for completing: By mark of the following personal qualifications.

The degree to which he is able to discriminate & evaluate facts to arrive at logical conclusions.

The degree to which his appearance and behavior cause people to react favorably.

The degree to which he is able to carry out orders with consistency & firmness to achieve objectives

Section IV. Check List

Directions for completing: From the following statements, determine whether or not each

statement applies to the officer under consideration. If a statement does apply, mark space one (1); if it does not, mark space two (2).

Applies	Does not Apply	
<u>1</u>	<u>2</u>	This officer has a broad and detailed knowledge of his profession
<u>1</u>	<u>2</u>	This officer's usefulness is limited to a narrow field
<u>1</u>	<u>2</u>	This officer does an excellent job of planning and organizing his work

B. SAMPLES OF ITEMS FROM SECTIONS OF THE OFFICER'S PROGRESS REPORT

Rating Scales

Indicate rating by check mark	Unsatisfactory	Fair	Good	Very Good	Excellent
Judgment	<u> </u>	<u> </u>	<u> </u>	<u> </u>	<u> </u>
General professional knowledge	<u> </u>	<u> </u>	<u> </u>	<u> </u>	<u> </u>
Proficiency in assigned duties	<u> </u>	<u> </u>	<u> </u>	<u> </u>	<u> </u>
Tact	<u> </u>	<u> </u>	<u> </u>	<u> </u>	<u> </u>
General fitness for the service	<u> </u>	<u> </u>	<u> </u>	<u> </u>	<u> </u>

Questions Eliciting Narrative Comments

Are you satisfied to have this officer? Yes ☐ No ☐ Give reasons

Handicaps

What are your recommendations for this officer's improvement?

Remarks

REFERENCES

- ADKINS, DOROTHY C. *Construction and analysis of achievement tests*. Washington, D.C.: U. S. Government Printing Office, 1947.
- APPEL, V., & KIPNIS, D. The use of levels of confidence in item analysis. *J. appl. Psychol.*, 1954, 38, 256-259.
- DWYER, P. S. The relative efficacy and economy of various test selection methods. *Psychol. Res. Section Rep.* No. 957, Adjutant General's Office, 1952.
- GARRETT, H. E. *Statistics in psychology and education*. New York: Longmans, Green, 1947.
- HARRIS, F. J., HOWELL, M. A., & NEWMAN, S. H. Forced choice tetrads—effect of scoring procedure and key length on validity and reliability. *Educ. & psychol. Measmt.*, 16, 454-464.
- MCMENAR, Q. *Psychological statistics*. New York: Wiley, 1949.
- NEWMAN, S. H. The officer selection and evaluation program of the U. S. Public Health Service. *Amer. J. publ. Hlth*, 1951, 41, 1395-1402.
- NEWMAN, S. H. Quantitative analysis of verbal evaluations. *J. appl. Psychol.*, 1954, 38, 293-296.
- NEWMAN, S. H., BUSSEY, R., & EPSTEIN, M. Performance criteria and evaluation methods for professional health personnel. Unpublished manuscript, 1955.

10. Sisson, D. E. Forced-choice—The new army rating. *Personn. Psychol.*, 1948, 1, 365-381.
11. Thurstone, L. L. Attitudes can be measured. *Amer. J. Sociol.*, 1928, 33, 529-554.
12. Witsell, E. J. The new officer efficiency report. *The Reserve Officer*, 1947, 24, 8-10.
13. Major study of comparative validity of five periodic officer efficiency reporting methods. *Personn. Res. Section Rep.* No. 670, Adjutant General's Office, 1945.
14. Studies of Officer Efficiency Report, WD AGO Form 67-1, in operation. I. Revalidation. *Personn. Res. Section Rep.* No. 791, Adjutant General's Office, 1949.

(Accepted for publication February 17, 1957)

Psychological Monographs: General and Applied

Impulse Expression as a Variable of Personality

NEVITT SANFORD, HAROLD WEBSTER, AND MERVIN FREEDMAN

Mary Conover Mellon Foundation, Vassar College

I. INTRODUCTION

WHEN one interviews large numbers of female college freshmen with a view to their educational needs, one finds it easy—perhaps all too easy—to divide them into two groups. On the one hand, there are those who seem primarily to need awakening, broadening, opening to experience; and, on the other hand, those who seem primarily to need self-discipline or organization or integration. The first group seem to have arrived at a structuring of personality which is perhaps premature and brittle, and one wonders whether much change can occur without experiences that might be shattering; the other group seem to have attained more freedom than they are quite entitled to, there being a serious question as to whether so much expression of impulse is under adequate control by the ego, to say nothing of the possibility that some of these latter subjects are “impulse ridden.” To speak somewhat more strictly, the impression is strong, when one studies these freshmen, that a dimension of considerable importance is that very broad one which extends from compulsiveness, authoritarianism, submissiveness, and passivity at one extreme to impulsiveness, irrepressibility, assertiveness, and adventurousness at the other.

In a research program that is focused upon changes in personality during the college years, we included measures of

trends that seem more or less to parallel this broad dimension. Thus, accenting compulsiveness, control, and the like, we have used the *Social Responsibility* (“Re”) scale of the California Psychological Inventory (12), the F scale (1), the K and Pt scales of the Minnesota Multiphasic Personality Inventory (“MMPI”) (13), and the *Fundamentalist Belief* items of Barron’s Inventory of Personal Philosophy (2); and with the accent on impulsiveness, expression, and the like, we have used the *Delinquency* (“De”) scale of the California Psychological Inventory (“CPI”), the D and Ma scales of the MMPI, Maslow’s scale for measuring *Dominance Feeling in Women* (19, 20), and the *Disbelief* items of the Inventory of Personal Philosophy.

As was to be expected, the scales within each of these two groups correlate positively with each other, and negatively with the scales of the other group, in our samples of college women. More than this, these scales have made sense when used to predict external criteria. Thus, students who withdraw from college in the freshman year are higher on De and lower on Re than are students who remain. And seniors are higher than freshmen on De, Dominance, Feeling, Disbelief, lower than freshmen on Re, F, Fundamentalist Belief. How shall we define what it is that the scales representing the two extremes of our broad dimension have in common? It has seemed to us that the compulsive-controlled extreme has been better understood and better measured. Our own “ideology-free” scale for measuring authoritarianism in personality (26) helps to cover the ground in this area. But obviously, “freedom from compulsiveness” or “lowness on the F scale” covers an enormous amount of territory that seems not to

have been very fully explored. Measures of Delinquency, of Depression, of Mania seem rather inappropriate in a study of college students. Although these scales have been of value, their distributions in our samples have been markedly skewed, with consequent loss of discrimination. Dominance Feeling and Disbelief seem aimed in the right direction, but they could hardly be expected to cover all the ground that would appear to lie there.

It seemed to us that there was need for an instrument that would measure a general readiness to express, to seek gratification of, impulses, in overt action or in conscious feeling and attitude. This readiness, though not so extreme as to produce high scores on the Delinquency or Mania scales, might nevertheless discriminate those subjects who, as noted above, seem to be in need of self-discipline and integration; and it might prove useful in further research.

II. THE IMPULSE-EXPRESSION SCALE

The first step in constructing the impulse-expression scale (hereinafter called for convenience the J scale) was to choose, from among the 677 true-false items in the test battery being used, those items which seemed most expressive of a readiness to gratify impulses directly in action. This procedure yielded 79 items that seemed clearly to express such tendencies as aggressiveness, rebelliousness, defiance, competitiveness, restlessness, excitability, adventurousness, unconventionality, sensuality, exhibitionism, tolerance, permissiveness, flexibility. The 79-item test was first scored for 237 seniors, and the correlations of each of the 677 items with this test score were obtained by a short method (25). The 123 items having correlations at the .001 significance level with the 79-item test, and which also had variances exceeding .08, were selected to comprise a second approximation to the final test; 77 of the

TABLE 1
MEANS, STANDARD DEVIATIONS, AND KR-21
RELIABILITY COEFFICIENTS FOR THE J
SCALE FOR THREE SAMPLES OF WOMEN

Group	N	\bar{X}	S	r_{tt}
Seniors	164	48.60	14.24	.862
Freshmen	220	41.43	13.66	.860
Alumnæ	50	38.96	12.96	.848

79 items from the initial criterion test had positive correlations, although only 49 of these (which also had sufficient variance) reached the .001 level of significance.

The 123-item test was next scored for two new samples, 164 seniors and 220 freshmen, and item-test correlations were obtained for each. Differences between the two sets of item correlations for these two samples were small and could easily be attributed to chance. Nevertheless, the 7 items for which the discrepancies were greatest were marked to be discarded subsequently, mainly as an added precaution to ensure that the test should measure the same thing for both freshmen and seniors; all retained items correlated at the .001 level of significance in the combined sample. The mean item-test correlation for the 123-item test for these two new samples was .32.

Next the 123-item test was scored for 50 alumnæ of the same college whose ages ranged from 38 to 46. Statistics for this sample, and for the 164 seniors and the 200 freshmen, are given in Table 1. The first sample, 237 seniors, could not be used for estimating reliability because it had been used for selecting the items.

The two largest mean differences of Table 1 are both easily significant. The fact that seniors possess more Impulse Expression than either freshmen or middle-aged women will be used later as an argument for the validity of the scale.

The difference between the freshmen and alumnæ means is not significant. The obtained distributions were quite symmetrical, but appeared to be slightly platykurtic. The reliability coefficients of Table 1 are lower-bound values, and KR-20 would have given slightly higher (but still lower-bound) values; KR-20 was computed for seniors only and found to be .886. Considering the diversity of the samples, the reliability appeared to be high enough, but the final revision of the test probably increased it slightly.

Statisticians differ somewhat in recommending how best to use internal consistency as a criterion for test construction. Ferguson (7) warns against using it as the only criterion. Cronbach and others (5, 16, 17) have discussed properties of the most popular measures of internal consistency, the Kuder-Richardson coefficients (14), which provide intuitively good indices of the unitary character of what is measured by a test. Whether factor tests measure single traits better than do tests which have high Kuder-Richardson coefficients, but may nevertheless be multifactorial, is still an open question; there appears to be no convincing reason why a single psychological trait, as it occurs in personality, should not be multifactorial. On the other hand, increasing homogeneity by item selection, even though insufficient as a condition to ensure a single factor test, usually will produce a test having high first factor saturations for most of the items. According to the criterion of internal consistency the items to be presented below comprise a good measure of what we have chosen to call *Impulse Expression*.

The final revision of the test consisted of discarding the seven items referred to previously and adding eight new

items which had a mean item-test correlation of .43 for the Senior sample. The new items were obtained from a *Weltanschauung* inventory constructed independently by Richard Jung. They were added after a study of item clusters showed that they would contribute substantially to test content.

The 124 items comprising the final J scale are given below, arranged into clusters according to content. The clusters represent agreement among the authors on certain general tendencies tapped by the scale, and they are presented as an aid to its understanding. The items could, of course, be classified in several different ways. Murray's variables of personality (22) seem particularly well suited to the task. Not only is it easy to regard each item as primarily an expression of some one of Murray's needs or traits, but his scheme offers terms for stating the positive side of the case, that is, for characterizing what the subject is seeking or exhibiting, so that we are not reduced to indicating merely that there is "absence of—" or "freedom from—." Moreover, the scheme offers more or less neutral rubrics for phenomena which would otherwise have to be described in the language of psychopathology.

Previous work has shown that virtually all the needs and traits necessary to a classification of the present items tend to be positively correlated. Sanford (24), using data derived from ratings of the overt behavior of school children, performed a correlation analysis of these needs and traits, and educed several clusters which can be of service here. We may not only group items that appear to belong together, on the basis of their surface meaning, but we may be guided by knowledge of what has actually cohered in previous work.

In the following, under each major heading, we list the needs and traits which have been shown to be intercorrelated and which are expressed in the items. "T" or "F" in parentheses after each item indicates the direction of response, whether "true" or "false," of subjects scoring *high* on the scale.

A. Ascendancy (n Dominance, n Aggression, n Autonomy, n Acquisition, n Recognition, n Defendancy)

Sanford (24) thought that the major theme of this cluster was "aggressive self-seeking," and that a basic concern was with "raising the status of the self," the subjects who exhibited the pattern being "not concerned with maintaining socially approved standards of conduct."

n Dominance, n Recognition, n Aggression

- When I work on a committee I like to take charge of things. (T)
- I tend to ignore the feelings of others when accomplishing some end that is very important to me. (T)
- I would be uncomfortable if I accidentally went to a formal party in street clothes. (F)
- When someone talks against certain groups or nationalities, I always speak up against such talk even though it makes me unpopular. (T)
- I dominate many of the men I know of about my own age. (T)
- Many of my friends would probably be considered unconventional by other people. (T)
- I would be uncomfortable in anything other than fairly conventional dress. (F)
- I have often met people who were supposed to be experts who were no better than I. (T)
- I would like to hunt lions in Africa. (T)
- I dislike women who disregard the usual social or moral conventions. (F)

n Autonomy, n Aggression, n Acquisition

- I must admit I find it hard to work under strict rules and regulations. (T)
- I have often either broken rules (school, club, etc.) or inwardly rebelled against them. (T)
- I have often gone against my parents' wishes. (T)
- I have sometimes wanted to run away from home. (T)
- I have the wanderlust and am never happy unless I am roaming or traveling about. (T)
- As a youngster in school I used to give the teachers lots of trouble. (T)

In school I was sometimes sent to the principal for cutting up. (T)

At times I have been so entertained by the cleverness of a crook that I have hoped he would get by with it. (T)

During one period when I was a youngster I engaged in petty thievery. (T)

If I could get into a movie without paying and be sure I was not seen, I would probably do it. (T)

At times I have a strong urge to do something harmful or shocking. (T)

Once in a while I feel hate toward members of my family whom I usually love. (T)

n Defendancy—with n Blamescape, n Retention, and Projectivity

I feel that it is certainly best to keep my mouth shut when I'm in trouble. (T)

If several people find themselves in trouble, the best thing for them to do is to agree upon a story and stick to it. (T)

I think nearly anyone would tell a lie to keep out of trouble. (T)

It is a good thing to know people in the right places so you get traffic tags and such things taken care of. (T)

It is a good rule to accept nothing as certain or proved. (T)

I go out of my way to meet trouble rather than try to escape it. (T)

I would do almost anything on a dare. (T)

I have very few quarrels with members of my family. (F)

One of Sanford's clusters was called "Aggressive Self-defense"; its dominant tendency was "to protect the self by taking the offensive, to overcome doubts or misgivings by adapting a somewhat hard and defiant attitude." It embraced all the variables listed above under Ascendancy except *n Dominance* and *n Recognition* and, in addition, *n Rejection*, *n Blamescape*, *n Retention*, and *Projectivity*. This pattern seems to be fairly well expressed in the last eight of the preceding items—those centering about *n Defendancy*. It turned out, however, that *Ascendancy* and *Aggressive Self-defense* were correlated .75, indicating a difficulty that will concern us much in the present work, namely, that whereas it is easy, and no doubt necessary, to separate in theory the "defensive"

from the "positive or spontaneous," it is very hard to do so at the level of overt expression. Moreover, "where there is a tendency to advance the self by energetic activity there is likely also to be present a tendency to defend it in the same way." It seems just as well here to make one grouping of *Ascendance* and *Aggressive Self-defense*.

B. *Sensation* (*n Sex, n Excitance, n Cognizance, n Recognition, n Succorance, n Exposition, n Blamescape, n Defenceance, Projectivity, Impulsivity, Emotionality, n Change*)

This is another syndrome from *Physique, Personality and Scholarship* (24): "a very large constellation which seems to denote a general tendency to seek for sensation or excitement." *Sensation* correlated .52 with *Ascendance*, and together they "seem to constitute a very general picture of 'outgoingness' . . . it would appear that in the *Ascendance* syndrome, *positive action* is the unifying tendency, while in the *Sensation* syndrome it is *positive feeling*." The *Sensation* syndrome was thought to embrace "a complex of which the main feature is the emotional expression of erotic needs."

n Sex, n Exposition, n Cognizance

- I like to talk about sex. (T)
- I like to hear risque stories. (T)
- I never attend a sexy show if I can avoid it. (F)
- I have never indulged in any unusual sex practices. (F)
- I have been in trouble one or more times because of my sex behavior. (T)
- When a man is with a woman he is usually thinking about things related to her sex. (T)
- I like to flirt. (T)
- In general I prefer the company of men to women (in sports, intellectual activities, hiking, theatre, conversation, etc.). (T)
- I am embarrassed by dirty stories. (F)

n Sex, n Excitance—with n Dominance and Masochism

- I have been disappointed in love. (T)

- I like men of whom I am a bit afraid. (T)
- I rather like justifiable conceit in a man. (T)
- I dislike a man who is frequently blunt in his speech. (F)
- I like men who antagonize me somewhat. (T)
- I prefer men who are never profane. (F)
- I rather like aloofness in a person I have just met. (T)
- I like worldliness in people. (T)
- I enjoy the company of strong-willed people. (T)

n Excitance

- I enjoy playing cards for money. (T)
- I enjoy betting on horse races. (T)
- When I get bored I like to stir up some excitement. (T)
- I have never done anything dangerous for the thrill of it. (F)
- I am fascinated by fire. (T)
- I like to go to parties and other affairs where there is lots of loud fun. (T)
- I think I would like to drive a racing car. (T)
- I have never done any heavy drinking. (F)
- I have used alcohol excessively. (T)

n Exhibition

- I like dramatics. (T)
- I would enjoy fame (not mere notoriety). (T)
- I would like to be an actor on the stage or in the movies. (T)

n Change—with Disjunctivity

- I have periods of such great restlessness that I cannot sit long in a chair. (T)
- I find it hard to keep my mind on a task or job. (T)
- I enjoy discarding the old and accepting the new. (T)
- I have several times had a change of heart about my life work. (T)
- I find that a well-ordered mode of life with regular hours is congenial to my temperament. (F)
- I would rather be a steady and dependable worker than a brilliant but unstable one. (F)
- I am known as a hard and steady worker. (F)
- I always see to it that my work is carefully planned and organized. (F)
- I am very careful about my manner of dress. (F)

Impulsion

- I often act on the spur of the moment without stopping to think. (T)
- I often do whatever makes me feel cheerful here and now, even at the cost of some distant goal. (T)

Emotionality

- I get excited very easily. (T)
- Once a week or oftener I become very excited. (T)

I easily become impatient with people. (T)
 I have had periods when I felt so full of pep that sleep did not seem necessary for days at a time. (T)
 At times I feel like picking a fist fight with someone. (T)
 At times I feel like swearing. (T)
 Sometimes I feel like smashing things. (T)
 Often I cannot understand why I'm so cross and grouchy. (T)
 I often feel as though I had done something wrong or wicked. (T)
 Sometimes I feel that I am about to go to pieces. (T)
 I frequently notice my hand shakes when I try to do something. (T)
 I work under a great deal of tension. (T)
 Something exciting will almost always pull me out of it when I am feeling low. (T)

One subgroup of items in the present set—those which have been labeled *n Sex* and *n Excitance*—deserve a special word. These items are adapted from Maslow's scale for *Dominance Feeling in Women*, and our label does not do them justice. It is clear that the items express *n Dominance*, *n Aggression*, and *Masochism*, as well as *n Sex* and *n Excitance*. As far as we know, however, Maslow's scale is the best verbal predictor of sex experience on the part of college women that has yet been devised. It owes its success to the fact that it comes to grips with some of the major psychological conditions of such experience—impulse, anxiety, effort to suppress anxiety by action, and dominant and provocative behavior toward men in order to elicit from them a sexual approach that is sufficiently dominant and aggressive to relieve anxiety and to gratify impulse.

As predictors of overt sexual behavior Maslow's items seem not out of place in the *Sensation* grouping. At the same time, however, they draw attention to the same question that was raised in connection with *Ascendance*, i.e., to what extent are we dealing with the straightforward gratification of impulse and to what extent with defensive acting out?

Sanford's "Anxious Emotional Expressiveness" syndrome, embracing *n Exhibition*, *Projectivity*, *Impulsion*, and *Emotionality*—which appear in the *Sensation* grouping—and in addition *Anxiety* and *Disjunctivity*, was said by him to have "much in common with the *Sensation* syndrome—the correlation is .73. . . . In both syndromes there is outgoing impulsive and emotionally toned activity, but this one is associated with anxiety, which gives to the behavior a quality of insistence or desperation. It might be that we have here an extraverted reaction to anxiety." It is possible to detect a note of insistence or desperation in many of the present items, but we do not have items that express anxiety or defensiveness in a pure way. Perhaps this is not too much to be regretted, for on the above evidence, such items would correlate highly with the others of the group.

C. *Endocathection and Intraception*

Endocathection (as opposed to *Exocathection*), according to Murray (22) has to do with the relative importance to the subject of "fantasy, reflection, imagination or abstract thought," and "practical, concrete, physical or social action." *Intraception* "is the disposition to be determined by diffuse personal feelings and inclinations (intangible subjective facts)," as opposed to *Extraception*, "the tendency to be determined by concrete, clearly observable, physical conditions (tangible, objective facts)." The two variables are easily confused, and they tend to be positively correlated (.38 in Sanford's study); nevertheless *Endocathection* may go with *Extraception*, e.g., when a subject is interested in ideas and theories about substantial events, and *Intraception* may go with *Exocathection*, e.g., when a subject tends

to live imaginatively, to dramatize the self, to express his sentiments and beliefs in action.

In the following items the accent is more on *Endocathection* than on *Intracception*, and there are intimations of *Projectivity* and *Narcissism*.

- I have had very peculiar and strange experiences. (T)
- I have had strange and peculiar thoughts. (T)
- My way of doing things is apt to be misunderstood by others. (T)
- I often get the feeling that I am not really part of the group I associate with and that I could separate from it with little discomfort or hardship. (T)
- I brood a great deal. (T)
- I have frequently found myself, when alone, pondering such abstract problems as free will, evil, etc. (T)
- I think I take primarily an aesthetic view of experience. (T)
- I dream frequently. (T)
- Sometimes without any reason or even when things are going wrong, I feel excitedly happy, "on top of the world." (T)
- A wise person thinks of life as a game; he is both in and out of the game and watching and wondering at it. (T)
- Some of my friends think that my ideas are impractical if not a bit wild. (T)
- I have had periods in which I carried on activities without knowing later what I had been doing. (T)
- I sometimes keep on at a thing until others lose their patience with me. (T)
- My home life was always happy. (F)
- I have had periods of days, weeks, or months when I couldn't take care of things because I couldn't get going. (T)
- Most nights I go to sleep without thoughts or ideas bothering me. (F)
- I often feel as if things were not real. (T)

D. Radical Sentiments

According to Murray (22), "the variable stands for the proportion of expressed sentiments, tastes, and opinions that are (1) novel, original, unique; or (2) contrary to those held by the majority of respected citizens." The sentiments of religious liberalism or of irreligiousness seem to fall well enough within the scope of this definition, though they might easily be taken as a group by themselves.

The other radical sentiments range over a wide area of human affairs.

- The only meaning to his existence is the one man gives himself. (T)
- God hears our prayers. (F)
- I go to church almost every week. (F)
- I believe in a life hereafter. (F)
- We cannot know for sure whether or not there is a God. (T)
- I pray several times every week. (F)
- I believe there is a God. (F)
- Organized religion, while sincere and constructive in its aims, is really an obstacle to human progress. (T)
- In religious matters I believe I would have to be called agnostic. (T)
- Moral codes are relevant only when they fit the specific situations; if the situations differ, they are merely abstract irrelevancies. (T)
- Man's search for a purpose, ideal, mission, etc., is largely a search for a plot or pattern to his life story—a story that is basically without meaning. (T)
- Children should receive no religious instruction of any kind. (T)
- People would be happier if sex experiences before marriage were taken for granted in both men and women. (T)
- Politically I am probably something of a radical. (T)
- I don't like modern art. (F)
- The unfinished and the imperfect often have greater appeal for me than the completed and the polished. (T)
- I think I am stricter about right and wrong than most people. (F)
- No man of character would ask his fiancée to have sexual intercourse with him before marriage. (F)
- I would disapprove of anyone's drinking to the point of intoxication at a party. (F)
- It is all right to get around the law if you don't actually break it. (T)
- I believe women ought to have as much sexual freedom as men. (T)
- In illegitimate pregnancies, abortion is in many cases the most reasonable alternative. (T)
- I would be ashamed not to use my privilege of voting. (F)

Although the items seem to fit rather well into the above categories, there is, of course, the usual difficulty with items that seem to express more than one idea and, hence, might almost as well be placed under some other heading than that to which they were assigned. And then, as mentioned above, there are

TABLE 2
CORRELATIONS OF J WITH OTHER SCALES

Scales Correlated with J	No. of Items	Items also in J	Alumnae Sample (N = 50)	Freshman Sample (N = 80) ^a
De, Delinquency	54	14	.40	.60
Re, Responsibility	41	6	-.60	-.27
MF II, Passivity	42	11	-.68	-.62
MF III, Intraception	33	5	.47	.35
MF II and MF III (Multiple R)	75	16	.76	.60
M Do, Maslow Dominance	47	17	.52	.62
E, Ethnocentrism		0	-.22	-.04
F, Authoritarianism		0	-.04	-.03
Vassar Developmental Scale, Developmental Status	124	37	.68	.51

^a A random sample of the freshman sample of Table 1.

other classification schemes. The items, after all, have been taken from existing scales—Delinquency, Dominance Feeling, Responsibility, Masculinity-Femininity and the like—and they might be classified on this basis. It would also be possible to classify them according to their most superficial meaning—rebelliousness, criticalness of authority, unconventionality, and the like. But however fascinating the classification exercise might be, *the most salient fact about the present group of items is their statistical homogeneity.* The first task is to conceptualize what it is that makes for unity in the scale. The scheme that we have adopted is instrumental to the performance of this task. We shall be in a better position to offer a theory, however, after we have considered some relations of the J scale to other measures.

III. RELATIONS OF THE J SCALE AND OTHER MEASURES

A. Correlations with Other Scales

Table 2 presents some correlations of J with certain other scales, scales which either overlap J or seem to stand in interesting relationships to it.

The *Delinquency* (De) scale from the California Psychological Inventory (CPI)

was designed by Gough "to indicate the potentiality for delinquent, troublesome behavior, and the tendency to rebel against authority and convention." It has with considerable success separated inmates of correctional institutions from other subjects. In view of Gough's expression of what his scale measures and of what has been said above about the J scale, a substantial correlation between the two instruments is to be expected.

It must be noted that 14 of the 54 De items are actually in the J scale. How much the correlations between J and other scales of Table 2 are inflated by common items is a bothersome question (the "common elements" correlation formula is inapplicable because of the presence of nonzero item intercorrelations) which we do not propose to answer here. If the overlap were removed, and the resulting *rs* then corrected for attenuation, it is doubtful that they would be much, if any, higher than the obtained *rs* of Table 2. Fortunately, nothing in our conclusions depends upon the absolute size of these correlations; but the number of common items in each case is given in Table 2.

The *Social Responsibility* (Re) scale of the CPI successfully identifies persons

who are "seen by others as responsible and dependable." It correlates around $-.45$ with De in large samples of college females. The correlations of Re and J, $-.27$ and $-.60$ (with 6 common items) seem to fit in well enough with our conception of high J as indicating a relative lack of concern with social obligations and standards of conduct.

The Masculinity-Femininity scales, MF II and MF III, require a few words of explanation. Included among our 677 true-false items were all the items from 3 MF scales: the CPI Femininity scale, the MMPI Mf scale, and a scale taken from the MMPI by Drake (6) which he standardized with the use of about 1,800 college undergraduates. We composed a first factor test, *f*, made up of items from these three scales, to which the factor gave weights of 5.0, 4.4 and 6.0 respectively. These weights were those obtained by extracting the first principal component, with unreliability allowed for. Since the reliability coefficients of the original tests were low (for the Drake scale, $KR-20 = .35$), it was not surprising to find that *f* had a reliability of only .61. Further efforts to maximize the reliability of *f* resulted in a coefficient of only .69. The test was therefore broken up into three rational clusters as follows:

MF I (18 items): Preference for conventionally feminine roles and interests.

MF II (45 items): Lack of aggressiveness, of dominance, of manipulativeness; docility, modesty, moral sensitivity.

MF III (33 items): Emotionality, fantasy, introspection, "neurotic trends," and aesthetic interests.

The reliabilities of these clusters were .43, .65 and .72, respectively. MF I is

represented in the J scale by a few items such as "I would like to hunt lions in Africa," but the scale is too short and unreliable to deserve serious consideration. After a study of the item-cluster correlations, three items were dropped from MF II, raising its reliability to .69. Further analysis to lengthen the clusters is in progress.

On the strength of its high correlations with MF II, $-.62$ and $-.68$ (with 11 common items), one might be inclined to say that the J scale contains a large element of masculinity. But the correlations with MF III, .47 and .35 (with 5 common items), suggest that J also contains a considerable element of femininity. There is a comment here on the old dilemma of how to define masculinity or femininity. We are inclined to doubt the fruitfulness of regarding femininity, in the sense of what, in attitude and in social behavior, differentiates women from men, as a single variable or factor. MF II and MF III seem to represent two quite different ways in which women manage the impulse life, by passivity and "masochism," or by elaboration of the "inner life." The correlations with J suggest that there might be a third way, by *impulse expression*, both in overt behavior and in conscious fantasy and attitude.

It is this, it would appear, that is expressed in the multiple correlations of J with MF II and MF III, .69 and .76 in the two samples. It will be interesting to know how men score on the J scale. However this might turn out, if we should be asked to consider how women differ psychologically from men, we should be inclined to ask in return, "Do you mean high-J women or low-J women?"

The Maslow scale (M Do) deals directly with one way in which women manage their femininity. It is not con-

cerned with sex differences; many of its items are not appropriate for men. As noted above, this scale taps a complex of factors, a syndrome one might say, that has much in common with J. The correlations, .62 and .52 (with 17 of the 47 revised Maslow items appearing in the J scale), are much as one would expect.

The absence of correlation between J and the E and F scales of *The Authoritarian Personality* is particularly interesting. One might suppose that a scale that embraced radical sentiments, endocathection, rebelliousness, and sensation-seeking would certainly correlate negatively with ethnocentrism and authoritarianism. Indeed, one of the types of low scorers on E and F distinguished by the authors of *The Authoritarian Personality* was the "impulse ridden." And there has grown up a common stereotype of the "low F personality," one that accents fluidity, lack of inhibition, sensuality, Bohemian values, and the like. The present results offer little support for this notion. Though such trends may be found in low scorers on F, they are by no means distinguishing features of such subjects. One of the reasons why the J scale does not correlate with F, we would suggest, is that the new instrument embodies too much aggressiveness. In a study of authoritarianism in relation to psychopathology (9) the present authors showed that impunitiveness is the low F scorer's most characteristic mode of handling aggression. (The F scale correlated $-.41$ with the MMPI Hy scale minus its somatic items.) Note, in this connection, the absence from the J scale of items having to do with positive social feeling; there is no *n Affiliation* nor *n Nurturance*, and little *n Succorance*. Unless a subject has some amount of these tendencies, it is probably difficult for him to be free of

authoritarianism and ethnocentrism. The same might be said for conscientiousness. The authors of *The Authoritarian Personality* have stated that comfortably internalized superego is the best guarantee against authoritarianism. If this is true, then high scorers on the J scale could hardly be expected to score very low in the F scale. Nothing could be more out of place in the J scale than expressions of a highly integrated superego. On the other hand, there are expressions of a primitive, unconscious agency that gives rise to anxiety and a need for punishment.

Since anti-intracception is a prominent feature of the authoritarian pattern, it might be supposed that *Endocathection* and *Intracception*, as expressed in the J scale, would go with low F scores. That this may not be so is food for some thought about authoritarianism. *Intracception* (being determined by diffuse personal feelings and inclinations), *Endocathection* (being devoted to general ideas, symbols and artistic productions), and *Radical Sentiments* do not in themselves stand in opposition to authoritarianism.

The Vassar Developmental Scale (VDS) is made up of 124 items, from our original battery of 677, which have differentiated freshmen from seniors at the .05 level in successive samples. The correlations of this scale with J, .51 and .68 (with 37 common items), are consistent with the fact, presented in Table 1, that seniors score higher on J than do freshmen. The suggestion is that something like a lifting of repression or a freeing of impulse occurs characteristically during the college years. The alumnae, it may be recalled, while not significantly different from the freshmen, were definitely lower on J than were the seniors. Let us hope that this reflects, on the part of the alum-

TABLE 3
MMPI MEANS FOR GROUPS OF FRESHMEN ARRANGED IN ORDER FROM
HIGH TO LOW SCORES ON THE J SCALE

Group	N	L	F	K	Hs	Hs ^a	D	Hy	Pd	Pd ^a
H	20	2.9	7.8	13.4	8.4	15.1	22.6	22.3	19.2	24.6
HM	44	3.4	6.7	14.7	6.5	13.9	21.0	22.5	16.7	22.6
LM	44	5.1	4.2	19.1	4.2	13.8	18.2	21.9	13.2	20.8
L	20	5.9	4.2	19.2	3.3	12.9	17.6	22.6	13.0	20.7
Group	N	Mf	Pa	Pl	Pl ^a	Sc	Sc ^a	Ma	Ma ^a	
H	20	34.4	12.0	20.9	34.3	22.0	35.4	20.9	23.6	
HM	44	34.7	10.2	16.4	31.1	15.2	29.9	19.6	22.5	
LM	44	35.5	9.0	7.7	26.8	6.7	25.8	13.8	17.6	
L	20	35.6	9.0	6.0	25.2	7.2	26.4	13.0	16.8	

^a Means in these columns have the "K correction."

næ, increased control rather than a loss of freedom!

Finally, some nonverbal correlates of J are only now being investigated, but one finding which seems worth commenting on has appeared. Subjects in two samples ($N_1 = 71$, $N_2 = 220$ freshmen of Table 1) were presented with a large number of figures, or designs, and asked simply to indicate whether or not they liked each figure. The figures varied according to three dimensions: simplicity-complexity, passivity-activity, and masculinity-femininity. The correlation of J with preference for active (as opposed to passive) figures was positive in both samples. The correlation with preference for complexity (or simplicity) or for masculinity (or femininity) of the figures was negligible, but, for example, in the second sample the correlation with a preference for active figures was .42. Further studies of this difference between high and low scorers on J are under way.

B. Relations Between Impulse Expression and the MMPI

The relations of J to the MMPI were studied in the sample of 220 freshmen. First the subjects were divided into 5 groups on the basis of their J-scale scores: the 20 highest (H), the 44 next

highest (HM), the 92 middle (M), the 44 who were low middle (LM), and the 20 lowest (L). This division of a distribution is recommended as most efficient by Flanagan (8). Then the mean scores on the various MMPI scales were calculated for each group, except for M. These means are shown in Table 3. Each row of the table may be regarded as a composite MMPI profile for each group of subjects.

The high and the low J groups display rather distinctive MMPI trends. Except for the *Hysteria* scale and, to a limited extent, the *Hypochondriasis* and *Masculinity-Femininity* scales, the high groups score consistently higher than the low groups on all scales, with the expected reversal of this tendency on the L and K scales. Although the L subjects are not sharply distinguishable from the LM's, a definite linear trend is evident for the four groups in the scales, with the H group scoring highest on the clinical scales (lowest on the L and K scales), and the HM group scoring intermediate between the H and LM groups.

The results shown in Table 3 suggest a significant relationship between J and some MMPI index of repressive-expressive tendency. One such index in common usage is the following: $Z \neq$

TABLE 4
DISTRIBUTION ON MMPI SCALES OF J SCALE ITEMS

MMPI Scale	L	F	K	Hy	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma
No. of J Scale Items	2	3	6	0	10	4	8	5	3	4	10	11

$K + Hy - Pd - Ma$. The correlation between scores on this index and the J score is $-.61$. The two measures, in other words, have approximately a third of their variance in common. It is our impression, based on some consideration of individual cases, that a subject who is high on the MMPI is very likely to be high on J, but that subjects high on J may have various types of MMPI profiles, including a generally low one. It would be possible, by assigning regression weights to the MMPI scales, to obtain a $J \times MMPI$ correlation more extreme than $-.61$, but J would still have variance that was uniquely its own. Of the 124 J-scale items, 38 are from the MMPI. Table 4 indicates the number of J items in each MMPI scale. Since some MMPI items are scored on more than one scale, the number of items in the table exceeds 38. It will be seen that the common items are distributed over all of the MMPI scales except for the *Hypochondriasis* scale, the greatest concentration being in the *Depression*, *Psychopathic Deviate*, *Schizophrenia*, and *Mania* scales.

To sum up the relations between J and the MMPI: higher scores on the former tend to go with "expression" rather than "repression," with a stronger inclination toward "psychopathology," and with a relatively stronger tendency to employ "psychotic" (the right side of the MMPI profile) rather than "neurotic" mechanisms of defense. It has been suggested before that these terms are not very appropriate for describing differences among college students. Perhaps

these MMPI terms when applied to patients in hospitals and clinics refer to more extreme instances of what we have called ascendance, aggressive self-defense, anxious expressiveness, endocathexis, and intraception.

C. The J Scale in Relation to an Adjective Check List

Another source of pertinent data was a list of 100 adjectives which the freshmen had marked "true" or "false" to indicate whether or not they believed them to be descriptive of the "Self," the "Ideal Woman," the "Average Woman," the "Ideal Man." These adjectives were obtained from one of the earlier forms of the Interpersonal Check List of La Forge and Suczek (15). They have been found useful in previous studies, and there is evidence that when they are used in such fashion as the above they possess retest reliability (4, 15) sufficient for most purposes. Subjects described Self and Ideal Woman on one occasion, the two other concepts on another, the same list of adjectives being used four times.

The subjects' test papers were divided into 5 groups according to the size of the J score, just as in Table 3. Taking the 100 adjectives one at a time, the number of subjects, in each of the 5 groups, who answered "true" was obtained. And this for each of the concepts, "Self," "Ideal Woman," and so on. Table 5, which gives the results for the adjective "sarcastic," is illustrative of the 100 tables that were prepared. Reading across one may note for each J group, the Highs, the High Middles and so on, and

TABLE 5
FREQUENCIES OF "TRUE" RESPONSES IN FOUR PARTS OF THE J SCALE
DISTRIBUTION FOR THE ADJECTIVE "SARCASTIC"

J Group	H	HM	LM	L	2×H	2×L	2×H+HM	2×L+LM
N	20	44	44	20				
Self	14	35	11	5	28	10	63	21
Ideal woman	2	6	1	0	4	0	10	1
Average woman	5	7	5	3	10	6	17	11
Ideal man	9	8	2	0	18	0	26	2

for groupings in which the scores for the most extreme groups have been weighted 2, the number of times "sarcastic" is ascribed to the Self, to the Ideal Woman, to the Average Woman and to the Ideal Man. It is not surprising to note that the high J subjects think of themselves, and of the Ideal Man, as sarcastic much more often than do those low on J (14 and 5, 9 and 0).

But the content of the adjectives will be taken up later. The concern here is with the two kinds of correlations which were obtained with the use of data from the tables illustrated by Table 5. The first kind of correlation concerns the question of whether or not, or to what extent, subjects scoring high and low on J differ in the way they conceive of themselves, of the ideal woman, and so on. Table 6 gives the high J-low J correlation of the 100 pairs of scores (illustrated in the last two columns of Table 5) for each of these concepts. Note that 11,000 weighted adjective scores were used in the computation of these coefficients (adjectives used by the M group receive zero weights). There should be no question of their reliability. The correlations are, however, inflated by variations in the adjective variances, that is to say, the totals for all the J groups tend to rise and fall together, as one considers one adjective after another, depending on the general popularity of the adjective in question. This means

that in interpretations we cannot rely very heavily upon the absolute size of these correlations.

The main conclusion from Table 6 is simply that differences in J scores are accompanied by differences in the way the self and others are conceived. The fact that the correlations are higher for "ideal man" and "ideal woman" than for "average woman" and "self" may indicate that the former are mainly cultural stereotypes not readily influenced by psychodynamic factors.

The second kind of correlation obtained with the present data concerns the question of whether, and how, subjects high and low, respectively, on J differ with respect to the relations among their conceptions of self, of ideal woman, and so on. The correlations are shown in Table 7, those for the high-J subjects being the upper entries, and those for the low-J subjects being the lower entries, in each row.

Viewing the table as a whole one may see that subjects high on J show greater differentiation in their conceptions of self and others than do subjects low on

TABLE 6
CORRELATIONS (INTERGROUP) OF WEIGHTED
ADJECTIVE SCORE PAIRS FOR THE
GROUPS HIGH AND LOW ON J

Self	Ideal Woman	Average Woman	Ideal Man
.71	.98	.88	.95

TABLE 7
INTERCORRELATION (INTRAGROUP) OF WEIGHTED
ADJECTIVE SCORES FOR THE FOUR CONCEPTS FOR
THE HIGH J GROUP AND FOR THE
LOW J GROUP

	Ideal Woman	Avg. Woman	Ideal Man
Self			
{High J Group	.47	.71	.59
{Low J Group	.87	.87	.85
Ideal Woman			
{High J Group		.55	.89
{Low J Group		.79	.94
Avg. Woman			
{High J Group			.58
{Low J Group			.80

J. This fact is further demonstrated by comparing the first principal components (Hotelling factors, extracted after placing 1's in the main diagonal of the correlation matrix) for the two matrices, the one comprising the high-J and the other low-J entries of Table 7. For subjects high on impulse expression the first component absorbs only 72% of the variation, for subjects low, 89%.

Differences in the way subjects high and low on J describe the self and idealized persons are striking. The ideal-man-ideal-woman correlations are large for both groups (.89 and .94), showing a common tendency to discriminate them less than other pairs. The self-ideal-woman correlation is much higher (.87) for the low-J group than for the high group (.47); in fact the latter, unlike the former, think they are more like the average woman (.71) than like the ideal woman (.47). High-J subjects also seem to think they are more like the ideal man (.59) than the ideal woman (.47), a tendency which is reversed in the lows (.87 compared with .85).

One interpretation of these findings is that subjects high on J are less stable—or should* one say less rigid?—in their

identities than are the low subjects. It may also be commented that relatively low self-esteem, perhaps even self-contempt, goes with aggressive self-seeking and self-indulgence. It may well be that the former is one of the forces behind the latter. Or perhaps the two are mutually influential.

Turning now to the content of the adjectives, to qualitative differences between subjects high and low on J, we note considerable difference between the two groups in types of adjectives chosen, as Table 8 indicates. Table 8 presents for the conceptions of "Self," "Ideal Woman," "Average Woman," and "Ideal Man" those adjectives which differentiate a group of high-J subjects ($N = 64$) from a group of low-J subjects ($N = 64$) at the .05 level of significance. Since all adjectives were marked either true or false by all subjects for each concept, interchanging the headings "True" and "False" in Table 8 will give the more frequent direction of response for low scorers, as contrasted with that of high scorers.

As can be seen from Table 8, the self-concept of the high group emphasizes qualities of aggression, suspiciousness, and stubbornness; that of the low group, cooperativeness, affiliation, and nurturance. The concept of the "Ideal Woman" does not differ greatly for the two groups, but while the high group emphasizes qualities of skepticism and autonomy, the low emphasizes passivity and submissiveness. The high-scoring group sees the "Average Woman" as much like themselves. She has many aggressive, suspicious, and stubborn qualities, although nurturance and generosity are stressed slightly more than in the self-conception. The low-scoring group, on the other hand, considers the "Average Woman"

TABLE 8

ADJECTIVES^a DIFFERENTIATING A GROUP OF
OF LOW SCORES ON J₁ CATEGORIES
(The scoring direction is that of high J₁)

Self	
True	False
Blunt	Affectionate
Competitive	Always smiling and pleasant
Complaining	Big hearted
Confiding	Businesslike
Cool and calculating	Can't refuse to cooperate
Critical of others	Charitable
Dominating	Devoted follower
Easily fooled	Encouraging others
Forthright and critical	Enthusiastic follower
Frequently disappointed	Firm but just
Gloomy	Fond of everyone
Hard boiled when necessary	Forgives anything
Hard to convince	Gentle and reassuring
Has childlike trust	Good leader
Impatient	Guiding
Irritable	Humble
Occasionally jealous	Likes everybody
Outspoken	Makes a good impression
Resentful	Managing
Resents being bossed	Modest
Sarcastic	Obedient
Selfish	Plays it safe
Self-punishing	Respected
Skeptical	Responsible and dependable
Snobbish	Self-reliant and assertive
Sometimes disrespectful	Self-respecting
Straightforward and direct	Stern but fair
Suspicious	Tender and soft-hearted
Too easily influenced	Trusting
Too lenient with others	Uncritical of others
Touchy and easily hurt	Unselfish
Wants everyone's love	Warm

Ideal Woman

True	False
Forgives anything	Apologetic
Forthright and critical	Businesslike
Hard to convince	Humble
Occasionally jealous	Obedient
Proud	Plays it safe
Skeptical	Self-punishing
Sometimes disrespectful	Wants everyone's love
Tender and soft-hearted	

TABLE 8—(continued)

Average Woman

True	False
Asking for help	Aggressive
Complaining	Always smiling and pleasant
Cool and calculating	Apologetic
Critical of others	Big hearted
Easily fooled	Businesslike
Easily led	Encouraging others
Forgives anything	Fond of everyone
Frequently disappointed	Forceful
Hard to convince	Humble
Impatient	Independent
Irritable	Modest
Outspoken	Obedient
Oversympathetic	Responsible and dependable
Resentful	Self-reliant and assertive
Selfish	Unselfish
Self-satisfied	
Shrewd and practical	
Skeptical	
Snobbish	
Soft touch	
Sometimes disrespectful	
Straightforward and direct	
Submissive	
Suspicious	
Too easily influenced	
Touchy and easily hurt	
Wants to be led	

Ideal Man

True	False
Blunt	Always smiling and pleasant
Critical of others	Apologetic
Dominating	Businesslike
Flattering	Can't refuse to cooperate
Forthright and critical	Dependent
Gives freely	Fond of everyone
Hard to convince	Humble
Occasionally jealous	Likes everybody
Sarcastic	Modest
Skeptical	Obedient
Sometimes disrespectful	Plays it safe
Straightforward and direct	Resents being bossed
	Uncritical of others
	Wants everyone's love

^a The adjectives were used by the subjects themselves to describe their conceptions of "Self," "Ideal Woman," "Average Woman," and "Ideal Man."

^b N=64 in both groups.

to possess fewer "negative" qualities. They see her as more affiliative and submissive than does the high-scoring group. Lastly, the "Ideal Man" presents considerable contrast. The low-scoring group stresses affiliative qualities, the high group aggressive and cynical qualities.

Here again it is possible to see the rather "negative" self-picture of the high-J subjects—negative according to the common values of our culture—and the rather positive or pleasant self-picture of those low on J. And one may note considerable correspondence between the contents of these self-pictures and the content of the J scale itself. "Blunt, aggressive" is like *n Aggression*, *n Dominance*, and *n Defendance*; "skeptical, distrustful" is like *n Rejection*, *n Autonomy*, *n Defendance*—and so on.

The undoubted tendency to self-depreciation on the part of the subjects high on J is rendered less ominous by the fact that the subjects themselves apparently do not see their ascribed "negative" qualities as so bad; they think that ideally a woman ought to be skeptical, autonomous, and perhaps somewhat hard-boiled.

Both groups seem largely governed by supplementary projection in giving their conceptions of the "Average Woman" and of the "Ideal Man." That the high-J group see the ideal man as aggressive and cynical, like themselves, does not mean that they like men any less than do the low-J subjects; on the contrary, the J scale itself, and particularly the *n Dominance* items, showed that it was qualities like aggression and cynicism that high scorers find most attractive in men; such qualities seem to serve as stimuli for the exercise of the same tendencies in themselves.

D. *Adjective Descriptions by Psychologists*

In all of the above attempts to reveal the meaning of the J scale, scores on the new instrument have been related to other measures based on the subjects' self-ratings. It would now seem desirable to consider scale scores in relation to an external criterion.

The 50 alumnae discussed in connection with Table 1 took part in a 3-day personality assessment project during which they were subjected to a wide variety of procedures, including ratings by staff members. Among the rating procedures used was a large adjective check list. Each of 5 raters independently checked in a list of 600 adjectives those which applied to each of the subjects. About 80 adjectives were found to discriminate (using χ^2 at the .05 level) the 14 highest scorers on the J scale from the 14 lowest scorers.

Adjectives most descriptive of the high and of the low scorers are presented in parallel columns, in the order of their discriminating power, in Table 9. These describe sharply contrasting patterns of behavior. The most discriminating adjectives—impulsive, irrepressible, loud, full of pep, erratic, vigorous, restless, impatient, and the like—seem to give a fairly vivid picture of impulse expression in its objective aspects. And similarly for the low end of the J distribution—retiring, silent, reliable, inhibited, reserved, stolid, and the like. We have not so far discussed the meaning of low scores on the J scale. A study of the discriminating adjectives suggests that such scores may be obtained either by subjects who are anxious and constricted or by subjects who are placid and at peace with themselves.

It is interesting to note the adjectives

TABLE 9
ADJECTIVES CHECKED BY 5 ASSESSMENT RATERS
TO DESCRIBE 14 LOW AND 14
HIGH SCORERS ON J

High Scorers	Low Scorers
Big-boned	Retiring
Impulsive	Silent
Irrepressible	Reliable
Angular	Inhibited
Loud	Reserved
Tall	Stolid
Full of pep	Humorless
Erratic	Shy
Vigorous	Withdrawn
Restless	Plain
Impatient	Faint-voiced
Stylish	Meek
Uninhibited	Painstaking
Excitable	Quiet
Outgoing	Weary
Humorous	Placid
Pleasure-seeking	Respected
Argumentative	Reticent
Critical of others	Uncommunicative
Assertive	Thorough
Egotistical	Fearful
Charming	Apathetic
Long-legged	Mild
Able to give orders	Passive
Tactless	Anxious
Mischievous	Modest
Dressy	Calm
Frivolous	Wan
Immoderate	Dignified
Worldly	Patient
Handsome	Dutiful
Witty	Cautious
Easy-going	
Sophisticated	
Bony	
Frank	
Self-centered	
Curious	
Aggressive	
Demanding	
Hasty	
High-strung	
Modish	
Good natured	
Illogical	
Delightful	

descriptive of physical features that distinguish the high scorers on J—big-boned, angular, tall, long-legged, bony. They are enough to suggest the possibility of a constitutional difference between the high and the low scorers.

IV. SUMMARY AND INTERPRETATION

The major findings that have been

reported in this paper are the following.

1. The scale developed for the measurement of impulse expression, the J scale, is statistically homogeneous, as judged from item-total correlations and reliability coefficients, but its constituent items are readily classifiable under the headings: *Ascendance, Sensation, Endo-cathection and Intraception, Radical Sentiments*.

2. The scale clearly has to do both with impulse expression that is more or less spontaneous and free, and with impulse expression that is defensive or "driven." These two tendencies appear to suffuse the scale as a whole rather than being separately represented by different items.

3. Seniors score significantly higher on the scale than freshmen, and higher than alumnæ in the age range 40-45. The difference between freshmen and alumnæ is not significant.

4. The scale correlates positively with *Delinquency (De)*, *Dominance Feeling in Women (M Do)*, a measure of "femininity" that accents the "inner life" (MF III), preference for "activity" in a figure-preference test, and a measure of developmental status based on empirically determined differences between seniors and freshmen (VDS). It correlates negatively with *Social Responsibility (Re)* and with a measure of femininity that accents "passivity" (MF II). Correlations of the J scale with the E and F scales of *The Authoritarian Personality* did not differ significantly from zero.

5. Scores on the J scale are positively associated with MMPI indices of expression (rather than repression), degree of "psychopathology," and tendency to use "psychotic" mechanisms of defense.

6. Subjects scoring high and subjects scoring low on the J scale differ in the

way they use an adjective check list to describe themselves, the "Ideal Woman," the "Average Woman," and the "Ideal Man." The high scorers show more differentiation among these conceptions; they give a relatively "negative" description of themselves and see themselves as more like their ideal man than like their ideal woman; aggressiveness and skepticism stand out in their conceptions of themselves and of the ideal man.

7. When an adjective check list was used by 5 psychologists to describe the behavior of a group of 50 alumnae, many more adjectives than would be expected by chance were found to discriminate subjects scoring high from subjects scoring low on the J scale. The adjectives distinguishing the two groups of subjects formed two self-consistent pictures, the one including "impulsive," "irrepressible," "impatient," and the like, the other "reticent," "reserved," "responsible," and so on.

How should these results be interpreted? We are not prepared to offer a comprehensive theory. More research is needed, particularly intensive clinical studies of high and low scorers, and renewed efforts to separate the spontaneous from the defensive aspects of the tendency to impulse expression. We may indicate here, however, the directions of our thinking at the present time.

Freud's (10) paper on "Libidinal Types" seems never to have attracted much attention, perhaps because it was too sketchy. He merely indicated the possibility of distinguishing Compulsive, Erotic, and Narcissistic types on the basis of the major direction of libidinal attachment—whether to internal agencies of control, to external objects of love, or to the self—and he went on to stress the commonness of mixed types. Sanford (23) leaned heavily on this scheme in working out a typology of true criminals. He distinguished the Anti-Social (Compulsive), who were capable of acting according to some principle, such as loyalty to a dissident group; the Pre-Social (Erotic), whose criminality proceeded from

their positive involvements (they were "easily led") or from their negative involvements with the objects of their love; and the Asocial, who had no loyalties but were out "to take care of number one" (Narcissistic).

It appears that according to this scheme the "defensive" high scorers on the J scale would belong to the Narcissistic type, or that narcissism would be the main feature of their libidinal strivings. The more "spontaneous" high scorers would also have a strong measure of narcissism, but this would be mixed with Erotic and Compulsive components. Some of the low scorers, probably the most extreme ones, would be of the Compulsive type, while others would be primarily Erotic with Compulsive elements.

We may consider first the defensive high scorers. The hypothesis of central narcissism seems to be supported by what has been said before concerning the J scale's accent on self-seeking, self-indulgence, self-defense, and the absence from it of items expressive of moral concern or of positive feeling for other people. Sanford (23), in writing about the asocial criminals, distinguished between the manifestly "egotistical," who sought by exhibitionism and manipulativeness to get something from other people, and the "withdrawn narcissists," the lone wolves, whose lives were dominated by fantasies and private schemes. Tendencies of both these kinds would seem to be present in the subjects high on J, though in less extreme degree than in the criminals. One might suppose that there were disturbances in the subjects' early relations with objects, so that there was instability in their introjections and distrust in their relations with people, leading the subjects to strike out for themselves. Of course, striking out for oneself, by self-forwarding and self-indulging tendencies, is not likely to raise self-esteem, not only because the subject must sooner or later bear the guilt for such behavior but because the behavior, however successful in a practical way, does not influence the unconscious sources of the narcissistic wound.

Barron (3), in a clinical study of the religious belief systems of the 50 alumnae used in the present research, came to the conclusion "that disappointment in the father and anger against him was the psychodynamic force which led to an affirmation of atheism. The atheistic claim that there is no God is essentially a denial of benevolent supernal power, as well as a repudiation of infantile wishes for love and for a staff of strength external to oneself." Since irreligiousness is an important trend in the J scale, it may be fruitful to explore this line of interpretation here. What holds for irreligiousness may also hold for other Radical Sentiments and for other trends in the J scale.

The hypothesis of disappointment in or anger

with the father is particularly relevant to our findings concerning masculinity-femininity in relation to J. A natural way for a girl to manage such feelings toward the father, or to respond to his loss through death or desertion, is by introjection of him. That such introjection was fairly common among our high-J subjects would be one way of accounting for the substantial negative correlation between this scale and MF II (passivity), and for the emphasis on activity, mastery, and daring in the J scale itself. But identification based on hostility and fear could not be expected to lead to the establishment of stable institutions within the personality. High-J subjects, as suggested earlier, have relatively unstable or perhaps unformed ego identities; they are by no means content to be "like father" or like a man; rather, it seems, they want to carry on certain kinds of relations with men. The correlation between J and the Maslow Dominance items, the ascription by high-J subjects of aggression and cynicism to the "Ideal Man," and the contents in the J scale itself suggest that although she wants to punish and dominate men, she nevertheless wants them to be tough, like herself, so that her aggression will not lead to any catastrophic consequences.

Premature sex experiences might be a factor in producing some of the attitudes just described. To hypothesize such experience would be also to give due attention to the element of Sensation in the J scale, and to the element of Delinquency. Adolescent gratification at the genital level creates a pattern of gratification-seeking that is not to be given up easily, even though its maintenance requires strenuous efforts to keep conscience in abeyance.

Disappointment in, and anger with, the father or his surrogates, however, is considered here as an example of something more general. We are hypothesizing as a major source of high-J scores experiences leading to disturbances in introjection and in relations with primary objects, such that the subject was forced to look to her self—to its advancement and to its protection. Such experiences usually occur first in a subject's relations with the mother. Indeed it seems likely that girls who are led into such relations with the father as we have described have first to deal with a mother who is relatively cold and capricious, or weak and inconsequential, or inflexibly strict.

It is to very early relations with the mother that one ordinarily looks for the circumstances leading to the differentiation in Compulsive, Erotic, and Narcissistic types of libidinal attachments and to the beginnings of those mechanisms of defense which seem to accompany them, e.g., reaction formation in the Compulsive, repression in the Erotic, denial and distortion of reality in the Narcissistic. Our high-J subjects may

have learned very early to prefer either action against the real world or denial and distortion of it as mechanisms of defense. It seems unlikely that the traumata of their early childhoods were severe enough to generate strong psychotic potential, but they may have been important enough to determine the type of response to later crises. It might be that in many of these subjects a narcissistic wound occurred at a time when a fairly substantial ego structure had been formed, so that they could feel, consciously, let down, disillusioned, embittered, insecure, and could determine to retaliate, to get what was rightfully their own, to make some radical changes in the world, to see to it that they were not again caught off guard.

It cannot be argued that all subjects who suffer disturbances in their early object relations become fixated upon the sort of defenses that have just been described. Such disturbances are also commonly regarded as major sources of depression, and very little is known of what determines the choice of reaction. At any rate, we are hypothesizing that just as action and denial become the defensive high-J scorer's preferred reaction to disturbances in early object relations, so depression becomes preferred in similar circumstances by the Compulsives who are extreme low-J scorers.

Concerning the more spontaneous high-J scorers it might be suggested that the mother herself was of the high-J type and had a free, give-and-take relationship with the father. She was not so cold or capricious as to create disturbances in the subject's early relations with her; and identification with her occurred later. The subject strove for the same type of relationship with the father as the mother had, and this was not sternly opposed from either side. The father was free enough to welcome responsiveness from his daughter, but able to require that her impulsive aims be sublimated. In these circumstances a girl could acquire a durable, if mild, superego and a fair measure of social feeling; and enough spirit to resist by action, or by provocativeness, the frustrations of a prolonged adolescence.

How are we to interpret the fact that scores on the J scale seem to increase during the college years? Two hypotheses may be suggested, the one having to do with increases in the more spontaneous aspects of J, the other with circumstances making for increases in the defensive types of impulse-expression.

It may be supposed that the college experience tends to bring about some lifting of repression; and that it does this by offering means for learning control at the same time that it stirs up the impulse life. If we have made a case for self-seeking and defensiveness as major features of the scale, we may also make a case for its ex-

pression of "self-actualization" in Maslow's (21) sense of this term. Many items of the scale seem to express autonomy, realism, acceptance of self and others, enjoyment of sensory pleasures, and spontaneity.

Fromm (11) and Maslow (21) have written about "selfishness" and genuine self-love as opposites, in the sense that the former springs from a lack of the latter. We have followed this line of thought in some of our formulations above. Now, however, we would propose that self-actualization may as well grow out of a personality organization of which selfishness, or the kind of defensiveness that we have described, is an integral part. "Genuine self-love" would seem to involve no less "self" and no less libidinal attachment than selfishness; the difference between them is that the object of attachment is, in the former, more worthy of it, and that there is enough sureness of that object and hence enough freedom so that, as in the case of Captain Shotover, the self-love may spill over onto the outside world. In other words, we would still argue that high scorers on the J scale were primarily Narcissistic, rather than Compulsive or Erotic, even though they owed their highness to a preference for the more sensible items and showed other signs of maturity. Growth toward self-actualization may occur as well in people who started from a Narcissistic orientation as in those who started from other orientations. It may be suggested, indeed, that subjects of the Compulsive or Erotic types have to pass through a narcissistic phase on the road to maturity.

That many seniors are in a narcissistic phase, and that this is largely provoked by the situation in which they find themselves, is our second hypothesis to account for their relatively high J scores.

We certainly do not wish to say that the seniors as a group are more self-actualized, or freer in the best sense, than the alumnae; rather, we should say that they are striving for self-actualization. Many seniors are in a situation of having thrown off traditional values without having fully established others of their own, of having loosened long-standing inner controls at a time when new experiences have to be integrated, of having rejected old identities at the very time when important decisions have to be made. We should not be surprised, then, if their strivings for self-actualization have an aspect of insistence, that they tend to be rebellious rather than autonomous, dominating rather than self-assured, cynical rather than realistic, hungry for sensations rather than able to enjoy them in a relaxed way. Among the hypothesized determinants of highness of the J scale, then, is a particular kind of developmental crisis, to which some subjects are more susceptible than others,

and through which different individuals pass at different times.

Most of the hypotheses suggested here call for comprehensive studies of life history variables in relation to J, and for intensive clinical studies of individuals scoring at the extremes on the scale. At the same time, and perhaps more immediately, the need is for studies aimed at differentiation among high and among low scorers on J. We have already expressed our interest in the possibility that the former might be divided into the "free" and the defensive, the latter into the constricted and the placid—or, perhaps, the Compulsive and the Erotic. Criterion groups of these kinds will, of course, have to be set up on the basis of clinical studies and used in connection with item analysis of the J scale. We shall also be able to use our sample of 50 alumnae for this purpose.

Another possibility is to take advantage of the absence of correlation between J and the F scale. This means that we undoubtedly have in our college samples large numbers of subjects who are high on J and high on F, high on J but low on F, and so on. Since much is known about the F scale, studies of this kind should throw some light on the meaning of J. But we may also have here a means for making significant differentiations among extreme scorers on F. For example, the woman who is high on J and F might well be one of those phallic, sadomasochistic kinds of women, whom we see occasionally on the political scene, who project their masculinity onto leaders or generals to whom they then submit masochistically. High or moderately high scores on J but low scores on F, on the other hand, might turn out to be a good indication of genuine freedom.

Our most immediate concern is with development during the college years. We have freshmen who are high and freshmen who are low on J. How do the two groups change, if they do, under the impact of education, and what are the circumstances of such changes? To take it the other way around, did the seniors who are outstandingly high on J start out as freshmen already fairly high, or has there been a radical "breaking loose" from a pattern of an opposite kind? Have seniors who are low or middle on J already attained a good measure of emotional control, or does there remain a potential for a "lifting of repression" at some later time? The suggestion, made above, that high scores on J may reflect a developmental crisis that is common in college seniors, means that we may have a particularly good opportunity for investigating, in a "momentary situation" as it were, the dynamics of impulse expression.

Finally, the J scale as we see it measures, like the F scale, a pattern of attitudes that is inti-

mately tied to unconscious sources in the personality. For this reason we should expect it to correlate significantly with many other measures of performance and of personality functions. We should not, however, expect these correlations to

be very high, just because J, standing at the level of attitude, may be expressed in overt behavior in different ways and have different sources within the personality.

REFERENCES

1. ADORNO, T. W., FRENKEL-BRUNSWIK, ELSE, LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.
2. BARRON, F. *Inventory of personal philosophy*. Berkeley: Univ. of California Press, 1952.
3. BARRON, F. The crisis in belief. (Mimeographed). Berkeley: Inst. of Pers. Assessment and Res., Univ. of California, 1955.
4. BILLS, R. E., VANCE, E. L., & MCLEAN, O. S. An index of adjustment values. *J. consult. Psychol.*, 1951, 15, 257-261.
5. CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-334.
6. DRAKE, L. E. Differential sex responses to items of the MMPI. *J. appl. Psychol.*, 1953, 37, 46.
7. FERGUSON, L. W. *Personality measurement*. New York: McGraw-Hill, 1952.
8. FLANAGAN, J. C. The effectiveness of short methods for calculating correlation coefficients. *Psychol. Bull.*, 1952, 49, 342-348.
9. FREEDMAN, M., WEBSTER, H., & SANFORD, N. A study of authoritarianism and psychopathology. *J. Psychol.*, 1956, 41, 315-322.
10. FREUD, S. Libidinal types. In *Collected papers*, Vol. 5. London: Hogarth Press, 1950.
11. FROMM, E. *Man for himself*. New York: Rinehart, 1947.
12. GOUGH, H. C. A preliminary guide for the use and interpretation of the California Psychological Inventory. Berkeley: Res. bull. of the Inst. of Pers. Assessment and Res., 1954.
13. HATHAWAY, S. R., & MCKINLEY, I. C. *The Minnesota Multiphasic Personality Inventory* (Rev. ed.). New York: Psychological Corp., 1943.
14. KUDER, G. F., & RICHARDSON, M. W. The theory of estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.
15. LA FORGE, R., & SUCZEK, R. F. The interpersonal dimension of personality: III. An interpersonal check list. *J. Pers.*, 1955.
16. LOEVINGER, JANE, GLEESER, GOLDINE C., & DuBOIS, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, 18, 309-317.
17. LORD, F. M. Estimating test reliability. *Educ. psychol. Measmt*, 1955, 15, 325-336.
18. LORD, F. M. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, 1955, 20, 1-22.
19. MASLOW, A. H. A test for dominance-feeling (self-esteem) in college women. *J. soc. Psychol.*, 1940, 12, 255-270.
20. MASLOW, A. H. Self-esteem (dominance-feeling) and sexuality in women. *J. soc. Psychol.*, 1942, 16, 259-294.
21. MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.
22. MURRAY, H. A. *Explorations in personality*. New York: Oxford Univ. Press, 1938.
23. SANFORD, R. N. A psychoanalytic study of three types of criminals. *J. crim Psychopathol.*, 1943, 5, No. 1, 57-68.
24. SANFORD, R. N. et al. Physique, personality and scholarship. *Monogr. Soc. Res. Child Develpm.*, 1943, 8, No. 1 (Ser. No. 34).
25. WEBSTER, H. Maximizing test validity by item selection. *Psychometrika*, 1957, 21, 153-164.
26. WEBSTER, H., SANFORD, N., & FREEDMAN, M. A new instrument for studying authoritarianism in personality. *J. Psychol.*, 1955, 40, 73-84.

(Accepted for publication February 26, 1957)

Psychological Monographs: General and Applied

A Behavioral Census of a State Hospital Population¹

STANTON P. FJELD

With

HARRIETTE S. ATKINSON, RUBEL J. LUCERO, BILL T. MEYER, AND ALLAN RECHTSCHAFFEN
Fergus Falls State Hospital, Minnesota

A SURVEY of the literature reveals no systematic inquiry into the behavioral characteristics of an entire hospital population. Yet this behavior is of concern both to those attempting treatment and to society as a whole. The actions of commitment courts, county welfare boards, relatives, clinicians, and scientists are based at least in part on opinions about the behavior of these patients. These opinions may or may not be accurate. There is urgent need for objective study of the actual behavior of mental hospital populations, so that various therapeutic programs can be better geared to variation in patient abilities and needs.

In the past, scientists in this field have generally limited their investigations to a small segment of those exhibiting behavior pathology. The recently admitted patient has been the subject of most state hospital research. Various generalizations have been made, without ade-

quate research, about the behavior of continued-treatment patients.

This is a report of an objective longitudinal study of the behavioral characteristics of an entire state hospital population. The authors hope that the results of this study may assist hospital personnel in planning therapeutic regimens, and ultimately aid the public in better understanding the behavior of the mentally ill. A further expectation is that the study will be of value in supporting or weakening various hypotheses.

The large amount of data, the number of factors, and the complex interrelations make it imperative that this paper be a "bird's-eye view." It is hoped that further and more detailed analyses of the data will be made.

THE HOSPITAL

Fergus Falls State Hospital is a 2,000-bed mental hospital located on the outskirts of a city of 12,000. Of the 1,086 acres that comprise the hospital grounds, 725 are farmed. Almost none of the 455 authorized positions are vacant. An average of 60 separations occur per year. Approximately 309 of the employees are primarily concerned with patient therapy and ward administration. The others are employed primarily for other purposes, e.g., kitchen, supply, maintenance, engineering, etc. Many of these people also have a great deal of patient contact (for example, about 75 patients are working in the main kitchens).

The budget for the fiscal year 1 July 1954 to 1 July 1955 was \$2,044,824. No major construction costs are included in this representative figure.

The hospital has three integrated (team approach) programs treating the following types of patients: recent admissions; regressed, long-term, and recently lobotomized patients; and continent, ambulatory seniles. Other wards in the hospital are screened periodically to provide patients for the latter two programs.

SUBJECTS

The subjects were the entire resident

¹ We are indebted to Starke R. Hathaway for the original idea; to W. L. Patterson, Superintendent, and J. G. Freeman, Clinical Director, for authorizing and assisting in the project; to John R. Hawkinson for criticizing the manuscript; to Adeline Foss for compiling most of the data; and to the psychiatric aides and nurses who made the ratings.

We are especially indebted to Reuben J. Silver for his assistance in preparing the final form of the manuscript, and to William Schofield, without whose constant criticism, encouragement, and active participation the project would never have been completed.

This project was financed in part by a research grant from the Department of Public Welfare of the State of Minnesota.

population of the Fergus Falls State Hospital. Two surveys were made: one in the first week of April 1951 ($N = 1,932$); and another, two years later, in the third week of May 1953 ($N = 1,925$). Of the patients in the first study, 1,481 were included in the second. Patients absent from the hospital for the whole week for any reason were not included.

The patients are drawn from a predominantly rural area of Minnesota. The largest city has a population of 28,410. (There are five cities with a population of over 10,000 and 26 of over 2,500.) The principal occupation of this portion of the state is agriculture.

PROCEDURE

Data were gathered on a total of twelve behavioral and nine case-history items. The nine case-history items are: Diagnosis, Present Age, Sex, Marital Status, Religious Affiliation, Physically Debilitating Condition, Total Length of Hospitalization, Education, and Age at First Admission.

Eleven of the behavioral items were derived from ratings of the areas on the L-M Fergus Falls Behavior Rating Scale, hereafter referred to as the L-M Scale (2, 3). The twelfth is an average of the ratings on these areas. The eleven areas are: A, Work; B, Eating Behavior; C, Behavior Toward Other Patients; D, Behavior Toward Psychiatric Aides and Nurses; E, Behavior Toward Doctors, Social Workers, and Psychologists; F, Attitude to Electric or Insulin Therapy; G, Participation in Occupational Therapy and Recreational Therapy; H, Attention to Dress and Person; I, Psychomotor Activity; J, Speech; and K, Toilet Behavior.²

Both in 1951 and in 1953, each patient was rated independently by two people on the ward (psychiatric aides and/or nurses). If the raters disagreed by three or more points with respect to their ratings of a patient in a given area, the rating on this area was considered invalid. The average of the two ratings for the area was used as the best estimate of the patient's behavior. In addition, the mean for all the areas (a more reliable measure) was computed for each patient. The patient's behavior was rated for the week prior to the date of the rating. Two weeks following the initial notice, all ratings were completed.

² The scale is reproduced in its entirety in the Appendix.

This scale does not give information with regard to thought processes, except insofar as inferences can be made.

There are some differences between the data collected in 1951 and in 1953, arising from: (a) minor differences in definition; and (b) more thorough examination of the patient's record in 1953. No changes in definition or method occurred with respect to Present Age, Age at First Admission, Diagnosis, Sex, Marital Status, and Length of Hospitalization. Age at First Admission and Length of Hospitalization refer to Minnesota State Mental Hospitals. Diagnosis is the latest one made by the hospital.

Religion was determined more accurately in 1953. The differences in definition of educational level and physically debilitating factors will be discussed later.

Some comparisons are made between the hospital population and the population of the State of Minnesota as determined by the United States Census Bureau for the year 1950 (7). The 38 counties in the 1950 admission district of the Fergus Falls State Hospital were used to determine the characteristics of the normal population. The population of these counties was 593,325, of whom 263,787 were males and 239,538 were females. These figures include only age groups over 15, since hospital admissions are rarely below this age. The hospital population is included in the 1950 census figures used in this comparison. Therefore differences found between the census and hospital populations will have maximal interpretive significance.

Since these populations are biased in the direction of rural areas, caution must be used in applying the results of this study to predominantly urban hospital populations. The characteristics of a rural, Midwestern people, many of Scandinavian descent and Lutheran faith, may differ quantitatively from those of other groups. In some instances, a direct comparison between the hospital population and the normal population gives ambiguous results. The authors have reduced the ambiguity of these results by the use of certain corrections.

RESULTS

The number of patients rated on the first census was 1,932, of whom 998 were male and 934 were female. Between the time of the first census in April of 1951 and the second census³ in May of 1953, a total of 451 patients left the hospital. This figure does not include patients

³ Hereinafter, the 1951 behavioral census shall be referred to as *Census I*, and the 1953 census as *Census II*.

TABLE 1
PRESENT AGE—HOSPITAL POPULATION

Age	Census I		Total				
	Male (N=998)	Female (N=934)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)
	%	%	%	%	%	%	%
16-25	2.7	2.6	2.6	2.8	1.1	8.3	8.5
26-35	11.4	11.7	11.6	8.7	7.8	11.7	12.8
36-45	20.7	18.0	19.4	19.7	21.3	14.2	20.9
46-55	21.0	18.9	20.5	20.4	22.5	13.5	19.9
56-65	17.1	23.2	20.0	19.1	21.0	12.6	22.8
66-75	14.4	14.1	14.3	16.0	16.1	16.0	9.9
76-up	11.7	11.2	11.4	13.2	10.1	23.7	5.2
Undet.	.1	.3	.2	.1	.1		
Mean	53.63	54.10	53.85	55.27	54.96	65.33	49.72
Median	52.18	54.78	53.29	54.68	54.57	57.59	50.11

Note.—Means and medians were computed from data grouped into finer categories than those in this table. The more detailed tables are available upon request. Data for the permanent population are taken from Census II.

who entered the hospital subsequent to Census I and departed prior to Census II; it does include 182 deaths and 58 transfers that occurred during this time. At the time of Census II, a total of 211 patients rated on Census I were absent from the hospital on either provisional or final discharge. The number of patients rated on Census II was 1,925, of whom 991 were male and 934 were female.

Hereafter, the following terms will be used to describe patient groups: "admissions"—the 444 patients rated only on Census II; "discharges"—the 211 patients rated only on Census I; "permanent population"—the 1,481 patients rated on both behavioral censuses.

A discussion of the population of Ferigus Falls State Hospital will be based primarily on the data from Census I, except where a more accurate determination was made in Census II, or when comparisons are made between the two censuses.

Present Age

There was no special problem connected with

the determination of age. For the past 25 years or so it has been customary to collect the birth date. Occasionally, there is some question, especially with older patients, or patients who were admitted a number of years ago, as to their exact age. The age of the patients is determined to the nearest year (for example, a patient who is 60 years, six months, and one day is called 61).

Approximately 36 per cent of the hospital population on Census I is 61 years of age or older, as compared with 39 per cent on Census II. The mean age of the patients on the second census is greater by 1.42 years. This difference is reliable⁴ and indicates a trend toward aging of the hospital population. Discharged patients are, on the average, younger than the hospital population.

The fact that the age shift is significant in only a two-year period is of extreme interest in terms of the direction of therapy. If this trend continues, greater attention must be paid to the care of a population which differs markedly from the younger psychotic patient. The older person has interests different from those of the younger person, and must have different types of treatments and activities in which to participate. This also means that the hospital has a smaller number of patients who are capable of performing arduous physical tasks. Therefore an increased number of employed personnel must be available to carry on the routine work of the institution. In addition, the physical layout of

⁴ In this paper differences reported as reliable uniformly meet the 1% level of confidence.

TABLE 2
AGE AT FIRST ADMISSION

Age	Census I		Total				
	Male (N=998)	Female (N=934)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. ^a (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)
	%	%	%	%	%	%	%
16-25 ^a	20.7	18.8	19.8	20.0	22.0	13.7	12.3
26-35	25.2	28.0	26.5	25.0	28.4	13.8	25.6
36-45	17.7	19.4	18.4	18.5	19.8	14.2	19.0
46-55	13.5	14.8	14.1	14.2	14.6	12.8	18.0
56-65	9.7	8.2	9.0	7.8	7.2	9.7	16.6
66-75	6.8	5.0	6.0	7.1	4.5	15.5	6.6
76-up	6.2	5.3	5.8	7.2	3.4	20.1	1.9
Undet.	.2	.5	.4	.2	.1	.2	
Mean	41.75	40.52	41.15	41.94	38.71	52.14	43.72
Median	37.45	37.00	37.23	37.70	35.36	52.56	41.47

^a Patients admitted before the age of 16 constitute approximately 1% of this age group.

the wards must be different in order to accommodate patients whose needs and capabilities differ from those of a younger group.

Age at First Admission

There were no special problems in the determination of age at first admission. In very rare instances in older records the age was not mentioned for a short hospitalization in a Minnesota

State Mental Hospital. The data in Table 2 indicate that the greatest percentage of admissions (26.5%) occurs between the ages of 26 and 35 inclusive. The Census II hospital population shows a statistically reliable shift toward an older age at first admission (mean, Census I = 41.2; mean, Census II = 41.9). Patients admitted between the censuses (i.e., rated on Census II, but not on Census I), average 54.1 years of age on first admission. The "permanent" population, on the other hand, has a mean first admission age of 38.7 years.

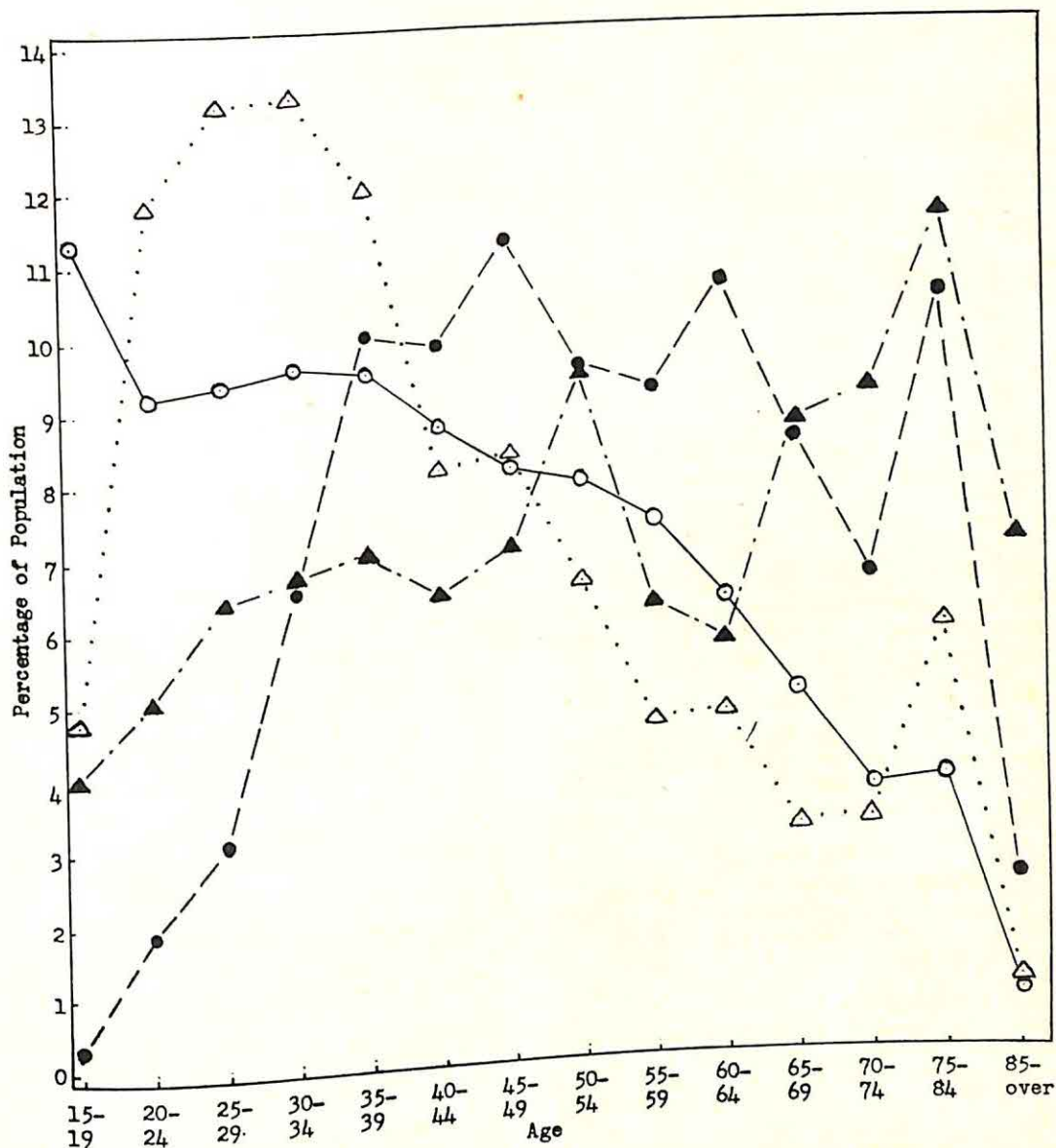
The comparatively high age at first admission of those admitted between censuses calls for comment. These are recent first admissions. Perhaps admissions to state hospitals are in fact older, reflecting the increased number of aged in the normal population. On the other hand, because of longer life expectancy, younger "chronic" admissions remain in the hospital longer than older "chronic" admissions and hence are more apt to be counted in a census. Perhaps younger admissions have poorer prognoses (for any type of discharge, including death) than do older admissions. These items are discussed later in this paper. Subtle changes, e.g., society's attitudes toward the aged, and changes in modern housing, may also be important in the increased admission age at this hospital.

Consecutive Admissions During One Year

Admission ages for a one-year admission period are given in Table 3. No special problems were encountered in this determination. During this one-year period the average age of read-

TABLE 3
AGE OF ADMISSIONS TO FERGUS FALLS
STATE HOSPITAL

Age	1 June 1925 through 31 May 1926	1 June 1951 through 31 May 1952		
	First adm. (N=428)	First adm. (N=375)	Readmissions (N=129)	
			Age at 1st adm.	Age at present adm.
	%	%	%	%
16-25	16.6	9.6	27.1	7.0
26-35	22.4	13.1	15.5	10.8
36-45	22.0	13.9	20.2	26.4
46-55	16.4	15.7	21.7	22.5
56-65	10.0	13.3	8.5	17.8
66-75	8.2	16.8	3.1	8.5
76-up	4.2	17.1	3.1	6.2
Undet.	.2	.5	.8	.8
Mean	42.43	53.91	39.37	48.91
Median	40.99	54.17	38.14	48.18



○—○ Present age of normal population 15 years or older. $N=503,325$.
 ●—● Present age of patients in the hospital on Census I. $N=1928$.
 △...△ Age at first admission of patients in the hospital on Census I. $N=1920$.
 ▲-.-▲ Age of first admissions from 1 June 1951 through 31 May 1952. $N=373$

FIG. 1. Present age of normal population compared with present age and age at first admission of hospital patients.

missions, at the time of readmission, was 49, and the average age of first admission was 54. This figure agrees substantially with the first admission age of the patients rated only on Census II (admission between censuses). The mean age of first admissions from 1 June 1925 through 31 May 1926 (42.4 years) agrees very closely with the age of admission of patients studied on

Census I (41.2 years).

Other data support the admission-age trends noted at the hospital. For example, the mean age of first admissions to all Minnesota state mental hospitals for the fiscal year 1953 was 53.8 years. In this hospital, first-admission figures for the years 1925 and 1951 show that the percentage of admissions over age 65 increased from

12.4% to 33.9% while the percentage over age 45 increased from 38.8% to 62.9%. Also, U. S. Census figures for 1920 and 1950 (7) indicate that the percentage of people over 65 (in Minnesota) increased from 4.7% to 9.0% while the percentage over 45 increased from 20.8% to 30.4%.

Age Comparisons

Figure 1 illustrates age curves for various populations. The Census I hospital population has a fairly constant percentage of patients between ages 35 and 70. On the other hand, the normal population has a steadily declining percentage of people in each age group from 20 through 85.

In the hospital population of Census I the majority of first admissions occur between the ages of 20 and 40. First admission ages for the hospital population cannot be compared directly with age groups in the normal population because the mean age of the normal population, at the time of admission of many of our patients (in the '20s, '30s, and '40s), was younger than it is now.

An estimate of the rate of admission from the age groups in the normal population (i.e., the 38 rural counties in the hospital receiving district) is best obtained by a comparison with the ages of consecutive first admissions to the hospital. During the one-year period 1 June 1951 through 31 May 1952, 504 patients were admitted to this hospital. Of these patients, 25.4% had previous admissions and were excluded. The 375 consecutive first admissions are, on the average, much older on admission than the patients rated on the hospital censuses. Considering the age distribution in the normal population, it appears that first admission to the mental hospital in the present study is, currently, most apt to occur when an individual is over 50 years of age.

The hospital has a lower percentage of people below the age of 35 than does the population-at-large (12% vs. 39.5%). From ages 45 to 85-and-over the hospital population has a relatively higher percentage than does the normal population (68.1% and 42.3%, respectively). Since the most frequent ages at first admission for the census population are 20 to 44, and since the patients have been in the hospital for some time, one would expect that they would distribute themselves in the present older age range.

Age-corrected samples are used in a comparison of items where age differentials must be considered. These age-corrections will be discussed later.

Marital Status

There were no important differences between Census I and Census II regarding the marital

TABLE 4
MARITAL STATUS

Status	Hospital Population					Census II Patient Sample (corrected for age) Compared with Normal Population			
	Census I		Total			Hospital Sample		Normal Population ^a	
	Male (N = 908) %	Female (N = 934) %	Census I (N = 1,932) %	Census II (N = 1,925) %	Perm. pop. (N = 1,481) %	Adm. betw. cens. (N = 444) %	Disch. betw. cens. (N = 211) %	Male (N = 270,280) %	Female (N = 245,796) %
Single	68.1	41.8	55.3	55.5	59.6	41.9	37.4	33.8	28.4
Married	21.1	38.3	29.5	28.8	26.0	35.3	45.5	61.1	63.9
Separated	.3	.6	.5	.7	.7	.7	1.0	5.1	7.7
Divorced	2.9	4.4	3.0	4.1	4.2	3.4	5.2		
Widowed	7.2	14.3	10.6	10.7	8.4	18.5	10.0		
Undet.	.4	.6	.5	.2	.2	.2			

^a Includes all age groups 14 years or over.

TABLE 5
RELIGIOUS PREFERENCE

RELIGIOUS PREFERENCE								
Religion	Hospital Population						Normal Population (N = 540,962)	
	Census II		Total					
	Male (N = 991)	Female (N = 934)	Census I (N = 1,932)	Census II (N = 1,925)	Perm. pop. (N = 1,481)	Adm. betw. cens. (N = 444)		Disch. betw. cens. (N = 211)
	%	%	%	%	%	%		%
Protestant	64.4	69.5	61.8	66.9	65.0	73.0	62.1	38.8
Catholic	24.4	24.7	24.6	24.6	25.2	22.5	25.1	26.0
Jewish	.1	.1	.1	.1	.1			3.1
Other	8.7	4.6	1.2	6.7	7.9	2.7	9.5	32.1
None	.2		8.7	.1	.2		.9	
Undet.	2.2	1.1	3.6	1.6	1.6	1.8	2.4	

status of patients. The differences between the normals and the Census I hospital population are all statistically reliable.

The hospital population has a much higher percentage of older people than the normal population. To render comparisons more meaningful, a correction was made for the age factor. These and all subsequent age-corrections were made by randomly selecting an appropriate number of patients from each age-group among the patients to approximate the age distribution of the normal population. The number of patients in the age-group 15 to 19 was too small to permit the drawing of an adequate sample. The actual frequency in this group is 13, while the theoretical frequency for a sample in this age-group would be 36. Except for this age-group, the sample is adequate.

On Census I, the majority (55%) of the patients are single, compared to only 28% of the normals. The age-corrected patient sample has a much higher proportion of single persons (69.8%) than either the normal or the uncorrected hospital populations. It may be that many patients do not marry because of seclusive personalities, passivity, the adverse reactions of others to them, or other unattractive biosocial factors. Some patients are removed from the environment before they reach marriageable age. On Census I, almost 15% of the hospital population is widowed or divorced, compared with only 8% of the normal population. In the age-only 8% of the normal population. In the age-corrected sample, the widowed-and-divorced category is equal to that of the normal population. Detailed 1950 census data for the entire state (7) indicate that 4.1% of males are widowed and 1.6% are divorced, while 10.6% of females are widowed and 1.9% are divorced. If these percentages apply to the population in the receiving district of the hospital, it is apparent that the hospital population has a much higher percentage of divorced people. (The preliminary U. S. Census figures for 1950 group the categories widowed and divorced. Detailed final figures are not available for the receiving district of the hospital.)

Religious Preference

Religious preference was determined from religion as given on the commitment paper or from the patient's own statement at the time of admission. There are instances in which the religious preference of the patient is not stated on the commitment papers, and there are also cases in which the patient is unable or unwilling at the time of admission to give his religious preference. Census II figures are quoted in Table 5 since this census has a more thorough check of the patient's record if religion was not listed on the admission note.

The percentages of church members of the predominant sects in Minnesota were taken from the 1936 Religious Census (6). These figures were determined from the membership rolls of the church, rather than by the questioning of individuals. Compared with the hospital population, more people in the normal population are classified as having no religion. The Catholic church membership as given agreed very closely with the percentage of Catholics found in the hospital. The percentage of Protestant church members outside the hospital was lower than was the percentage in the hospital. In the community a Protestant church member will be dropped from the church rolls if his attendance is irregular, since the church must pay a certain sum per member into the national organization. Many people may not be registered as members, but, if asked, would state that they were. In the Roman Catholic church, a person who has been a Catholic is regarded as such until his defection becomes "public knowledge," whether or not the member is active.

Because of these differences in determination of religion the percentage figures for normal population and hospital population must not be compared directly. Catholic percentages agree closely in the normal and hospital populations. It is likely that Protestant percentages are more in agreement than the unexplained data indicate.

TABLE 6
PHYSICAL DEBILITATION

Status	Hospital Population							Census II Patient Sample (corrected for age) Compared with Normal Population	
	Census I		Total						
	Male (N = 998)	Female (N = 934)	Census I (N = 1,932)	Census II (N = 1,025)	Perm. pop. (N = 1,481)	Adm. betw. cens. (N = 444)	Disch. betw. cens. (N = 211)	Hospital (N = 291)	Normal (N = 593,325)
	%	%	%	%	%	%	%	%	%
Without	85.0	84.6	84.8	85.6	88.6	75.4	92.4	89.4	99.1
With	15.0	15.4	15.2	14.4	11.4	24.6	7.6	10.6	.9

Physically Debilitating Conditions

Defining physical disability was difficult. In Census I all patients whose behavior was impaired because of physical factors were counted as debilitated. Chronicity of the illness was not considered. In Census II, people with temporary disability were not included. For this reason patients recovering from operations, broken limbs, or bedridden with acute short-term illnesses, were not counted as being physically handicapped. If the illness was severe enough to be noted on the physical examination, or if it was felt by a physician to be chronic rather than acute, or if the illness would permanently limit the behavior of the patient, it was classed as a physically debilitating factor. Amputation, blindness in one eye, or partial uncorrectible difficulty in seeing or hearing (provided these were serious enough to be mentioned in the notes on physical factors) were all counted as physical handicaps. Epileptics were not counted as handicapped unless their condition was so severe that it necessitated their being bedridden. Patients who suffered from a partial paralysis, or from a marked loss of a sense of balance were classed as physically handicapped. The figures for the normal population were derived from the number of beds in medical hospitals and in rest homes. This figure does not include all of the chronically disabled people in the community, many of whom remain in private homes. However, the disparity between those listed as disabled in the population at large and those in the hospital population seems too great to be accounted for wholly on this basis. In the age-corrected sample (previously used in Table 4) 10.7% of the patients are classified as debilitated. The disparity between the normal and hospital populations remains great.

The aged frequently have physical debilitation and personality change occurring in combination, and the high percentage of aged in the hospital population partially accounts for the high frequency of physical disability. In the younger patient the presence of a physical handicap may result in certain unfavorable changes in personality which necessitates institu-

tionalization in a mental hospital. People who have a physical handicap and an unstable personality tend to remain in the hospital rather than attempt a marginal adjustment on the outside. Perhaps all of these factors contribute to the relatively large number of physically handicapped found in the hospital.

Length of Hospitalization

The data for Table 7 were taken from Census I. It should be noted that the bulk of the population is made up of so-called chronic, long-term patients, and that, despite the relatively large number of admissions which this hospital receives in the course of a year, only 14% of the population has been here less than two years. This figure includes approximately 100 recent admissions.

The mean length of hospitalization of the census population is 11.22 years. The patients who stay two years or more (86% of our population) present the greatest therapeutic challenge. Even though present treatment methods are inadequate for them they accounted for 35% of discharges between censuses. The number of patients hospitalized between six and twenty years tends to remain relatively constant. After this time a sharp decline is found, which probably reflects the increasing death rate of those patients. Also of interest is the difference in mean length of hospitalization between males and females. This difference, which is reliable, may reflect the slightly longer life span of the female, since females are approximately the same age as males on first admission. Perhaps, also, females are sicker or slower to recover than males.

Educational Level

Educational level in Table 8 is based primarily on Census II since definitions of the levels agree more closely with those of the U. S. Census Bureau. All "common school" educations are placed in the 5-to-7-year group. (In Census I "common school" educations were regarded as equivalent to graduation from the eighth grade, since only six years of schooling were available in many areas of this state or in foreign coun-

TABLE 7
LENGTH OF HOSPITALIZATION

Years	Census I		Total				
	Male (N=998)	Female (N=934)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)
	%	%	%	%	%	%	%
.5 or less	7.4	3.8	5.7	6.7	.1	28.8	26.5
1	8.0	8.4	8.2	8.3	.2	35.4	21.8
2	9.5	8.4	9.0	5.8	1.5	20.1	16.1
3	5.7	3.2	4.5	5.2	5.4	4.7	7.1
4	5.7	5.4	5.5	5.0	5.8	2.5	5.2
5	3.4	4.6	4.0	4.1	4.9	1.4	3.3
6-10	15.7	18.9	17.3	16.9	20.9	3.4	9.0
11-15	20.1	15.5	17.9	15.6	10.9	1.4	7.1
16-20	11.8	13.4	12.6	14.6	18.6	1.4	2.4
21-25	5.4	7.2	6.3	7.9	10.1	.2	1.4
26-30	2.9	4.8	3.8	4.2	5.3	.5	
31-35	2.9	2.4	2.6	2.8	3.6	.2	
36-up	1.4	3.2	2.3	2.6	3.4		
Undet.		.8	.4	.3	.3		
Mean	10.48	12.02	11.22	11.71	14.77	2.72	3.72
Median	8.75	9.80	9.30	9.91	13.31	1.10	1.60

tries.) The term "specialized education" includes college as well as the usual varieties of training which, although not customarily given grade designations, have a more or less formal sequence of tasks to be mastered. On-the-job training usually was not regarded as specialized education unless accompanied by formal classroom work. The designation "specialized education" does not necessarily imply graduation from high school. Completion of the eighth grade is sufficient for inclusion, since it was not uncommon for an individual to have graduated from the eighth grade and then to have taken one or two years of training in a specialized subject such

as agriculture. For the present generation graduation from high school is generally a prerequisite to further training. All those who have received any form of college, nurses' training, business school, etc., are grouped in the category "specialized education" and are not listed under any other category.

The percentage of undetermined educations is relatively high and constitutes 2.4% of the hospital population (Census I=10.7%). This may reflect the reluctance of relatives to state education when the patient has completed only a very few years of school or when they feel that his school performance has been unsatisfactory.

TABLE 8
EDUCATIONAL LEVEL

EDUCATIONAL LEVEL								Census II Patient Sample (corrected for age) Compared with Normal Population	
Level	Hospital Population								
	Census II		Total						
	Male (N=991)	Female (N=934)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)	Hospital (N=249)	Normal (N=503,325)
	%	%	%	%	%	%	%	%	%
0-4 grade	23.6	18.8	14.3	21.3	22.6	16.9	9.5	19.7	7.0
5-7 grade	31.6	27.3	18.0	29.5	31.1	24.1	17.5	26.5	17.7
8th grade	26.4	23.4	37.0	24.9	24.1	27.5	40.8	30.9	37.9
9-12 grade	11.6	17.3	13.9	14.4	13.2	18.5	21.3	18.9	25.0
Any spec. education			6.1	7.5	7.1	9.0	6.6	3.2 ^a	10.8 ^a
Undet.	4.4	10.8	10.7	2.4	1.9	4.0	4.3	.8	1.6
	2.4	2.4							

^a College only.

The percentage of unknowns is also partially explained by the antiquity of some records.

Comparison with the educational level of the general population presented a problem. U. S. Census estimates for 1947 indicate that 68% of the population in the age-group 30 to 35 had completed more than eight school grades, contrasted with only 28% age 65 and over. The disproportionately high percentage of older people tends to lower the educational level of the hospital population, necessitating the use of age-corrected data in making comparisons with the normal population. Table 8 provides data on the educational level of the age-corrected sample.

In order to render the data comparable with the U. S. Census, the term "specialized education" in the age-corrected sample was redefined as college training only, and excludes business schools, nurses' training, etc. As a result of this redefinition, the percentage in this classification declined from 5.5% to 3.2%. Most of those excluded are now in the group with 9 to 12 years of schooling.

It is apparent that a high percentage of the hospital population has a poorer education than the normal population. Although the feeble-minded in the hospital population undoubtedly influence data on educational level, the percentage so diagnosed (8.6%) appears too small to materially alter the discussion. The achievement of the hospital population remains considerably below that of the normal population even after excluding from consideration all

diagnoses of mental deficiency. Perhaps the personality handicaps of some patients are reflected in an inability to function efficiently or to remain at a task to completion (e.g., completing school). In our society low intelligence (and/or achievement) is usually associated with low income, as well as other social disadvantages, which tend to impose a psychological burden. Perhaps psychoses are associated with low intelligence in our society, or perhaps the more intelligent may, through their extra resources, or resources of friends and relatives, less often go to a state institution, and perhaps the more intelligent have a better prognosis and are less apt to remain in a state institution.

Diagnostic Classification

Diagnoses given in this paper (Table 20) are based on the revised 1934 statistical classification systems (4). The hospital population follows the national trend. Approximately 50% of the patients have a diagnosis of schizophrenia. A diagnosis of mental deficiency is given 8.6% of the patients, versus an expectation, on the basis of psychological tests, of a maximum of 2½% in the general population. Of these mentally deficient patients, 32.3% are diagnosed as being without psychosis and 67.7% are diagnosed as with psychosis. The high percentage of patients diagnosed as without psychosis is partially due to the long waiting list at the state school for the feeble-minded. On the basis of educational level most diagnoses of mental deficiency appear reasonably accurate.

TABLE 9
ADMISSION DIAGNOSES—I JUNE 1951 THROUGH 31 MAY 1952
(N = 504)

Code	Diagnosis	N	Code	Diagnosis	N
000	Undiagnosed	5	17	Manic-depressive psychoses	36
01	Psy. with syph. meningo-enceph.	4	171	Manic type	20
02	Psy. with other syph. of CNS	2	172	Depressed type	10
05	Psy. due to alcohol	2	174	Mixed type	4
07	Psy. due to trauma	1	18	Schizophrenia	121
08	Psy. with cerebral arterioscl.	51	180	Unclassified	2
09	Psy. with disturb. of circul.	10	181	Simple type	22
10	Psy. due to conv. disord. (epil.)	4	182	Hebephrenic type	5
11	Senile psychoses	83	183	Catatonic type	6
111	Simple deterioration	51	184	Paranoid type	59
115	Paranoid type	15	185	Other types	27
12	Involuntional psychoses	62	19	Paranoia and paranoid cond.	6
121	Melancholia	39	20	Psy. with psychopath. personal.	1
122	Paranoid type	13	21	Psy. with mental deficiency	10
13	Psy. due to metabolic, etc., dis.	1	22	Undiagnosed psychoses	11
15	Psy. due to unknown or heredit. assoc. with organic change	8	23	Without mental disorder	44
16	Psychoneuroses	41	230	Unclassified	2
161	Hysteria	10	232	Alcoholism	23
165	Reactive depression	16	234	Mental deficiency	5
166	Anxiety state	3	236	Psychopathic personality	8
			237	Other non-psy. disease	6
			24	Primary behavior disorder	1

Admission Diagnoses

Table 9 includes diagnoses of all admissions⁵ between 1 June 1951 and 1 June 1952 and permits comparisons with the frequency of diagnoses found among the patients in our census. The proportions of admission diagnoses of psychoses with cerebral arteriosclerosis, psychoses with other disturbances of circulation, senile psychoses, involutional psychoses, psychoneuroses, manic-depressive psychoses, and alcoholism are considerably higher than the proportions among patients included in the census. Conversely, the proportions of the diagnoses of psychoses due to unknown or hereditary cause but associated with organic change, schizophrenia, psychoses with mental deficiency, and certain disorders without psychosis (excluding alcoholism) are higher among the census population. The above trends probably have some relationship to the prognoses of patients in these categories.

Later discussion will take into consideration the shifts in diagnoses that have occurred between the censuses, in an effort to determine the effects of treatment, discharge, and death.

Behavioral Ratings

The appendix shows each of the areas rated on the scale. The number preceding each statement does not appear on the scale. Only a blank space is provided for a check mark. Absolute interval equality is not claimed for statements that are made under each one of the areas rated, but it is felt that each of the statements represents a behavioral advance over the previous one. Needless to say, not all patients were rated on all areas, because some patients were not observed in a particular form of behavior; also, some ratings for an area were discarded as invalid. The average percentage of undetermined ratings on areas "A" through "K" (excluding "F") is 4.3%.

Area "A," Work

The average female does a little work with a lot of urging and needs constant supervision. On the other hand, the average male is much more apt to have a regularly assigned job under supervision. (These statements are the best verbal approximations of the behavioral ratings and assume face validity of scale statements.) This difference between the sexes is both statistically reliable and clinically significant. There is no reliable difference between the two censuses.

It can be seen from Table 10 that males are

⁵ The term "admissions" in this paragraph refers to admissions for this one-year period and not to "admissions" as defined earlier in the study.

much more capable of functioning at the level of a rating between 3 and 5 points in the area "Work." A patient rated 4 is performing at a relatively normal level. In most such cases, only a minimum of supervision is required, and the patient is able to make the routine decisions that the job necessitates. Twenty-one per cent of the men and 17% of the women perform at this level.

An explanation of the superiority of males is difficult. Perhaps the majority of jobs in the hospital are of a type with which males are more familiar. Males are more frequently employed in a large establishment where functions such as farming, laundry, cafeteria, construction and maintenance, and janitor work closely approximate those of the hospital. Much of the housework of the institution bears little, if any, resemblance to that done by the average housewife. Dishwashing, for example, is a matter of perhaps 10,000 dishes per meal and is done on an assembly-line basis. Perhaps the male worker in the hospital enjoys greater prestige and receives more consistent rewards. Since the need for male workers is much greater, hospital industries may be more willing to devote time to training and keeping them. There are more open men's wards (in which patients have freedom of the hospital grounds) than open women's wards. It is also possible that women patients as a group are somewhat more variable in their behavior.

Area "B," Response to Meals

Males and females are approximately equal. The average patient seems to eat by himself using a knife, fork, and spoon properly, but may at times have some problem. It would also seem that the average patient will ask for things to be passed at the table but will do little else. Very few patients require special attention at meals. On Census I a total of only 159 patients, or 8% of the rated population, required this form of attention. There were no reliable changes between the first and second censuses.

Area "C," Attitude to Other Patients

Males are reliably better in this regard than females. On Census I the average female is somewhat unfriendly but may speak to others and even have a friend. The males, on the other hand, are more consistently rated as being friendly and having some spontaneity in making contacts and initiating play or work of a fairly high order. Approximately one-third of the patients are quite seclusive and/or hostile. About one-half are quite friendly and operate at a fairly high social level.

Patient contact is of a higher order than is generally assumed. The greatest difference between males and females is in the range 4 to 5.

TABLE 10
RATINGS OF PATIENTS ON THE AREAS OF THE L-M SCALE

Behavior Area and Group	Behavior Rating					Mean		
	1 %	2 %	3 %	4 %	5 %	Census I	Census II	Dis- charges
A. Patient's attitude to work								
Male ^a	38.7	11.4	27.2	8.4	12.4	2.48	2.50	3.46
Female ^b	47.8	16.9	17.9	7.3	9.4	2.20	2.14	3.21
Total ^c	43.1	14.1	22.7	7.8	11.0	2.34	2.32	3.34
B. Patient's attitude to meals								
Male	7.3	22.8	32.3	24.0	13.3	3.22	3.20	4.05
Female	9.2	29.4	29.2	10.1	21.7	3.17	3.14	4.22
Total	8.2	25.9	30.8	17.3	17.4	3.20	3.17	4.13
C. Patient's attitude to other pa- tients								
Male	31.7	14.1	32.9	11.0	8.7	2.56	2.65	3.67
Female	36.6	17.6	32.9	5.9	6.6	2.35	2.42	3.42
Total	34.1	15.8	32.9	8.5	7.7	2.46	2.53	3.55
D. Patient's attitude toward psychi- atric aides and nurses								
Male	20.5	26.7	24.8	15.5	8.8	2.74	2.87	3.81
Female	22.6	35.4	18.3	8.5	14.0	2.67	2.59	3.87
Total	21.5	30.9	21.7	12.1	11.3	2.70	2.74	3.84
E. Patient's attitude toward doctors, social workers, and psychologists								
Male	4.6	25.5	41.7	13.1	5.2	2.94	2.90	3.61
Female	10.3	32.3	40.7	7.2	1.1	2.62	2.64	3.30
Total	7.4	28.8	41.2	10.2	3.2	2.79	2.78	3.47
F. Patient's response to electric or insulin therapy								
Male	1.0	1.3	4.6	1.0	.6	2.90	2.90	3.41
Female	3.5	6.7	6.4	1.0	.2	2.44	2.38	2.64
Total	2.2	3.9	5.5	1.0	.4	2.60	2.51	2.88
G. Patient's attitude to occupational and recreational therapy								
Male	34.1	17.7	18.1	6.9	9.4	2.36	2.53	3.56
Female	30.3	27.7	20.0	12.4	6.1	2.45	2.53	3.46
Total	32.2	22.5	19.1	9.6	7.8	2.40	2.53	3.52
H. Attention to dress and person								
Male	9.3	33.8	14.3	19.2	14.6	3.01	3.12	3.94
Female	14.9	32.9	21.8	10.8	18.2	2.94	2.83	4.00
Total	12.0	33.3	17.9	15.1	16.4	2.97	2.98	3.97

Note.—Because the rating for each area is the average of two raters, a patient may receive a rating that is not a whole number. (This occurs in approximately 10 per cent of the cases.) The percentages in a given row do not add up to 100 per cent because of a small proportion of "undetermined" ratings.

^a Cen. I, N=998; Cen. II, N=991; Disch., N=1113.

^b Cen. I, N=934; Cen. II, N=934; Disch., N=98.

^c Cen. I, N=1,932; Cen. II, N=1,925; Disch., N=211.

TABLE 10—(continued)

Behavior Area and Group	Behavior Rating					Mean		
	1 %	2 %	3 %	4 %	5 %	Census I	Census II	Dis- charges
I. Psychomotor activity	12.3	17.2	12.2	25.4	20.1	3.35	3.57	4.29
Male	19.6	15.9	16.5	20.5	24.2	3.26	3.28	4.31
Female								
Total	15.8	16.6	14.3	23.0	22.0	3.30	3.42	4.30
J. Speech	20.3	20.7	8.5	18.5	28.0	3.22	3.18	4.34
Male	18.8	14.0	14.8	22.5	26.0	3.35	3.12	4.47
Female								
Total	19.6	17.5	11.5	20.5	27.0	3.28	3.15	4.40
K. Toilet behavior	9.2	4.3	12.9	15.7	56.5	4.10	4.16	4.70
Male	11.4	7.6	22.4	16.0	39.1	3.73	3.91	4.58
Female								
Total	10.3	5.9	17.5	15.8	48.1	3.92	4.04	4.65

Although the difference is not statistically reliable both males and females improved slightly between Census I and Census II. If this trend continues it may be a reflection of the increased emphasis on socialization.

Area "D," Response to Psychiatric Aides and Nurses

Males and females do not differ reliably. The average patient in this area does most things when asked and may make simple requests. On Census II males showed an improvement which approached statistical reliability. Females declined in this area. Patients are not as hostile as one might assume. The results of increased social programming throughout the hospital may have paid greater dividends on the male side of the hospital. This finding is consistent with the results of a *total-push* program for repressed patients (1, 5).

Area "E," Response to Doctors, Social Workers, Psychologists

The lower number of patients rated in this area (90.8%) reflects the lack of contact with non-nursing professional personnel in the hospital. Again males are found more often at the upper levels. A statistically reliable sex difference was found on both censuses. Although the evidence thus far presented from the census indicates that male patients show better behavior than female patients, it must be remembered that the overlap is great. In this area, for example, 9% of the females have a higher rating than

80% of the males and 5% of the males have a lower rating than 89% of the females.

Area "F," Response to Electric or Insulin Therapy

This area concerns patient attitudes toward the administration of treatments, and only recently treated patients are rated. 251 patients were rated on Census I and 539 on Census II. On both censuses males were less antagonistic and apprehensive than females. There was no reliable difference between Census I and Census II in the average response to this form of treatment. The larger N on the second census undoubtedly represents an increased use of electric and insulin therapies in the hospital.

Area "G," Occupational and Recreational Therapy

On Census I males and females did not differ reliably, but females tended to be higher. This may have been due to the effective occupational and recreational therapy program that Fergus Falls has carried on for a long time, which was somewhat more extensive on the women's side of the hospital, and which emphasized things that women characteristically do, e.g., embroidering, fine painting, etc. In Census II it was found that the males had shown a statistically reliable improvement which made them equal to the females. The interim increase in recreational and occupational therapy programs on the male wards and the initiation of an aide training program (resulting in greater acceptance of, and more interest in activities) may explain patient improvement in this area.

TABLE II
AVERAGE OF AREAS "A" THROUGH "K"—"L"

Group	Census I grouped by average rating					Mean behavior rating				
	1.0—	2.0—	3.0—	4.0—	5.0	Census I	Census II	Perm. pop.	Admissions	Discharges
	1.9 %	2.9 %	3.9 %	4.9 %	%					
Male ^a	17.1	30.8	32.0	20.1		2.96	3.07	2.96	3.28	3.90
Female ^b	23.3	31.2	27.4	17.7	.3	2.82	2.84	2.76	3.18	3.84
Total ^c	20.1	30.9	27.8	18.9	.2	2.89	2.94	2.86	3.24	3.87

^a Cen. I, N=998; Cen. II, N=991; Perm. pop., N=735; Adm., N=256; Disch., N=113.

^b Cen. I, N=934; Cen. II, N=934; Perm. pop., N=746; Adm., N=188; Disch., N=98.

^c Cen. I, N=1,932; Cen. II, N=1,925; Perm. pop., N=1,481; Adm., N=444; Disch., N=211.

The average patient in occupational and recreational therapy will participate when asked, but will require frequent urging. He may show some spontaneity.

Area "H," Attention to Dress and Person

Although the difference is not reliable, the males tend to be slightly better than the females. The average patient has some interest in his appearance. A comparison between the first and second censuses shows that the males improved, whereas the females tended to decline. The difference between males and females on Census II is reliable. Scale artifact may be responsible for the somewhat atypical distribution of patients in the various behavioral levels of this area.

Area "I," Psychomotor Activity

The average patient exhibits some purposeful behavior and also some behavior which is a result of the illness. The males showed a highly reliable increase in their behavior rating on the second census (gain = .22). At this level activity is usually purposeful although movements may be somewhat erratic in terms of hypo- or hyperactivity. (There are more female than male hyperactive wards in the hospital.) As a result of the improvement in male behavior the hospital population was reliably better on Census II.

Area "J," Speech

There was no reliable difference in speech between males and females, although the females were slightly higher on the first census. The average in this area is quite high. Many of our patients speak in short, clear sentences if they have a request to make. There was a reliable decrease between Census I and Census II for the female population of the hospital (from 3.35 to 3.12). The males remained constant. This is the first statistically reliable fluctuation that the females have shown in the present study.

Area "K," Toilet Behavior

As a group, patients are relatively high in terms of toilet behavior. With some exceptions such as being too neat or spending too much time or perhaps being incontinent on rare occasions, the average patient is normal. Again the males are, on the average, superior to the females, and the difference is large enough to be both reliable and clinically significant. Both males and females improved on Census II, and the female and total populations were reliably higher. This improvement may reflect the discontinuance of the various physical restraints in the hospital, and the gradual retraining of patients who had been untidy and incontinent for a number of years.

It is generally conceded that women need more time and space in the bathroom in order to be "properly groomed." It should be mentioned that toilet facilities in the hospital are more adequate for men. For example, wards of comparable size have the same number of stools but the men's wards also have urinals.

"L," Average of Behavior Ratings of Patients

An average of the ratings of a patient on the eleven behavioral areas is more reliable and representative of the patient's total behavior than any one area of the scale. Minor fluctuations and variations will tend to be canceled out. The behavioral average of patients has been demonstrated to be both reliable and meaningful (3). Areas of the scale have not been studied so carefully.

The average for the entire hospital was slightly higher on Census II, but the difference was not statistically reliable. Males were reliably better than females on both censuses. Males also showed a statistically reliable improvement between Census I and Census II, while females showed

TABLE 12
AVERAGE BEHAVIOR RATINGS OF 50 STUDENT NURSES COMPARED
WITH THE HOSPITAL POPULATION (CENSUS I)

Group	Area									
	A	B	C	D	G	H	I	J	K	L
Nurses	4.6	4.7	4.0	4.4	4.2	4.9	4.8	4.8	4.9	4.6
Patients	2.3	3.2	2.5	2.7	2.4	3.0	3.3	3.3	3.9	2.9

essentially no difference. The population that was in the hospital on *both* censuses (the "permanent" population) also showed a reliable difference between males and females, the males being better. This very consistent tendency for males to be rated higher than females may be of considerable importance to the administration of a mental hospital. The permanent population shows no variation in average behavior ratings between Census I and Census II, although they had a lower behavioral average than those not rated on both censuses.

The difference between average behavior ratings for males and females, either admitted or discharged, indicates that females are approximately .1 lower than males. But male admissions are not reliably higher, nor male discharges reliably lower in behavioral rating than female admissions and discharges; these facts, then, do not seem to account for the improvement shown by the males between Census I and Census II.

It was found in a follow-up of a group of patients treated on a *total-push* program (1, 5) that males tended to improve more and hold their improvement while females improved less and did not hold their improvement. The Ferris Falls hospital population exhibits a clear, consistent tendency for mentally ill females to have a lower average behavior rating than mentally ill males.

These findings suggest a need for more staff on the female wards of the hospital. Reorientation of the hospital administration toward the problems of female patients may be of some assistance in improving general behavior. Perhaps more attention should be devoted to such things as make-up, the fitting of clothing, the types of jobs assigned to female patients, and so on.

The differences in behavior ratings found to exist between Census I and Census II cannot be explained with certainty at this time. Perhaps the initiation and maintenance of an expanded therapeutic program (dating from increased funds in 1949) resulted in a gradual improvement in the behavior of patients. It will be necessary to make further behavioral censuses at equal intervals to answer a number of questions regarding the stability and areas of improvement noted on these censuses.

Ratings on Student Nurses

One might suppose that the L-M Scale would not differentiate adequately among normals since most of them would receive a rating of approximately 5, which is the upper limit of the scale. Presumably, however, this average would give some indication of differences that may exist between "normals" and mental patients.

A group of 50 student nurses were rated by two of their classmates on the L-M Scale. The students lived together in a nurses' home for three months prior to the ratings, and had also worked on different wards in the hospital. The average rating that these students received was 4.6. Four students received an average rating of 4.2 and five received an average rating of 4.3. The two areas rated lowest were "C" (response to peers) and "G" (occupational and recreational therapy). The rating was approximately equal to that of the highest-level ward in the hospital. This ward has only long-term and convalescent patient workers (patients' average was 4.6). It should be remembered that students are young, single females, of above-average intelligence, who were rated on behavior in a less restricted environment than that of the patients.

These results imply that more subtle things than the overt behaviors measured by the scale play a role in adjustment. Perhaps attitudes toward others or toward hospital living are very important. Certainly many patients at the level of 4.5 could leave the hospital if they chose to do so (and some do). Some patients at this level are unable to interact satisfactorily with other people and are not permitted to leave.

The rankings of patients and students are given in Table 12. The groups show the greatest differences on Areas "A" (work), "G" (occupational and recreational therapy), "H" (attention to dress and person), "D" (response to immediate supervisors), and "C" (response to peers).

Such things as present age, age at first admission, and so on were studied in relation to the average behavior rating of patients, and these results are quoted and discussed below. Unless otherwise noted, these analyses are from Census I.

TABLE 13
BEHAVIOR RATINGS IN RELATION TO
DEBILITATING CONDITIONS

Status	Census I Population		Census II Patient Sample (corrected for age)	
	N ^a	M	N	M
Nondebilitated	1,638	3.04	260	3.04
Debilitated	294	2.38	31	3.13
Total	1,932	2.89	291	3.05

^a Includes approximately .5% undetermined cases.

Behavior Ratings of the Physically Debilitated

There were 294 patients (150 male, 144 female) who were considered to have physical handicaps. Since these handicaps would affect the ratings rather markedly, especially on such things as psychomotor activity, the authors excluded these people in the belief that such averages would be spuriously low.

In the non-age-corrected data there is a marked difference in average behavior-rating between those patients who are physically debilitated and those who are not. This finding would be expected since many of the areas on the behavior rating scale presume that the patient is physically able. The age-corrected sample of the hospital population was analyzed for average behavior rating of the physically debilitated. The exclusion of a large number of elderly patients in the age-correction process resulted in a substantial improvement in the average rating of the physically debilitated group. In this age-corrected sample the behavior rating of the debilitated is equal to that of the non-debilitated.

On items with statistically reliable differences between the debilitated and nondebilitated, the results will be discussed for the two groups separately.

TABLE 14
AGE AND LENGTH OF HOSPITALIZATION IN RELATION TO DEBILITATING CONDITIONS

Item (in years)	Non- debilitated (N = 1,638 ^a)	Debilitated (N = 294 ^a)
Mean Present Age	52.10	63.62
Median Present Age	51.18	66.50
Mean Age at 1st Adm.	39.21	51.98
Median Age at 1st Adm.	35.81	53.00
Mean Length of Hosp.	11.50	9.66
Median Length of Hosp.	10.03	5.83

^a Includes approximately .5% undetermined cases.

Behavior Ratings in Relation to Present Age

Nondebilitated patients between the ages of 51 and 65 have an average behavior rating slightly higher than that of any other age group. Patients under 40 and over 75 tend to show poorer behavior than the average patient. There is a modest curvilinear relationship ($\eta = .15$) between present age and behavior rating. The physically debilitated population does not exhibit a relationship between behavior rating and present age. Aged patients are more frequently classified as debilitated. The physically debilitated population is 10 years older than the total hospital population. In the Census II age-corrected sample, the physically debilitated patients are older (mean = 50.6) than the nondebilitated (mean = 43.5).

Because of some differences in the definition of debilitation, caution should be exercised in a comparison of Census I and Census II data.

Behavior Ratings in Relation to Age at First Admission

Table 15 indicates a curvilinear relationship between behavior rating and age at first admission ($\eta = .27$). Up to age 65, patients show a trend toward higher behavior rating as their

TABLE 15
AGE AND LENGTH OF HOSPITALIZATION IN RELATION TO BEHAVIOR RATINGS
OF NONDEBILITATED PATIENTS

Item	N	r	η (eta)	P.E. _{η}	χ^2	n	p
Present Age	1,635	+ .05	.15	.016	33.4	11	.001
Age at 1st Adm.	1,633	+ .10	.27	.016	110.0	12	.0000
Length of Hosp.	1,630	- .13	.19	.016	31.5	6	.0000

age at first admission increases. This increase in behavior rating is in accord with the hypothesis that the older an individual is upon becoming mentally ill, the less severe has been his degree of maladjustment in society. A priori, a patient who is admitted before the age of 25 would have more difficulty in adjustment than an individual admitted between 40 and 65. A person admitted beyond the age of 65 has more difficulty in hospital adjustment than a somewhat younger patient. On age at first admission, the physically debilitated population is thirteen years older (mean = 52.0) than the nondebilitated population (mean = 39.2). The relative percentage of physically debilitated is much higher in the older age groups. Physical deterioration (reflected in a lower behavior rating) seems to play an important role in the admission of the aged patient.

Behavior Ratings in Relation to Length of Hospitalization

Patients hospitalized two years or less have a definitely higher behavior rating than other patients. There is a steady regression in behavior rating up to 20 years of hospitalization. After this time the ratings begin to improve and those patients who have been hospitalized 26-45 years show behavior that is equal to or higher than the average patient in the hospital. After 45 years there is a very marked regression in behavior, probably as a result of aging. The most interesting aspect of this relationship is the definite cessation of regression that occurs. This suggests that behavioral regression is not continuously progressive. This hypothesis is supported by the highly reliable difference between the correlation ratio and the product-moment correlation (cf. Table 15).

Obviously, the hypothesis of continuing regression is inadequate. The hospital's experience with its older long-term patients has indicated that many show spontaneous remissions. The illnesses are frequently said to have "burned out." The needs of an individual of 50 years are, of course, different from those of a person of 30. Observation indicates that a considerable number of aged patients could be discharged if they were not so dependent upon the hospital. Within the institution they are capable of functioning at a very high level.

This phenomenon of improvement after 20 years of hospitalization emphasizes the importance of studying the long-term mentally ill. Generalizations made from short-term patients are not necessarily applicable to all mental illness.

Physically debilitated patients have a median length of hospitalization approximately four years less than that of the nondebilitated population. The median for debilitated males is

TABLE 16
BEHAVIOR RATINGS IN RELATION TO RELIGION

Religion	N	M
Protestant	1,008	3.07
Catholic	404	3.02
Other	26	3.00
No Religion	144	2.96
Total	1,582	

approximately four years and the median for females is approximately eight years. The general trend is consistent with that of the total hospital population, but the physically debilitated patients are older and their length of hospitalization less, probably because of a high mortality rate. The hypothesis that there is higher mortality among the physically debilitated, especially among males, is supported by their older age at first admission, and their very short length of hospitalization. If the physically debilitated were included, the improvement shown by patients after 20 years of hospitalization would be even more marked because of the lower behavior rating of physically debilitated patients and the shorter length of their hospitalization.

Age at first admission and length of hospitalization are inversely related. The relationship of present age to these two items should also be considered in interpreting census data. For example, the relation of present age to behavior rating is undoubtedly influenced by both the age at first admission and the length of hospitalization. Though present age may be the same, the patient who was the younger at first admission will tend to have the longer hospitalization and the lower behavior rating.

Behavior Ratings in Relation to Religion

There is no relationship between religious affiliation and mean behavior ratings. Patients who belong to minor religions or who have no religion do not differ markedly from those who belong to one of the major religious sects. In the normal population members of minority sects and those who profess no religion are frequently considered to be maladjusted.

Physically debilitated patients have essentially the same proportion of religious faiths as the nondebilitated population.

Behavior Ratings in Relation to Marital Status

The lower rating of widowed patients may reflect their advanced age. Single patients are lower in behavior rating than those who have been married or divorced. Divorced patients have

TABLE 20—(continued)

Code	Diagnosis	Male		Female		Total					
		Census I (N=998)	Census II (N=991)	Census I (N=934)	Census II (N=934)	Census I ^a ex. p.d. (N=1,603)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)
12	Involuntal psychoses N Mean	22 3.82	15 3.73	43 3.10	46 3.23	59 3.41	65 3.34	61 3.35	36 2.98	25 3.90	23 3.96
	121 Melancholia N Mean	14 4.20	8 4.13	18 3.31	22 3.22	29 3.76	32 3.70	30 3.46	16 2.91	14 4.09	17 4.02
	122 Paranoid type N Mean	7 3.34	6 3.48	24 2.98	21 3.15	28 3.14	31 3.06	27 3.22	20 3.03	7 3.76	5 4.17
15	Psy. due to unknown or he- reditary assoc. with organic change N Mean	13 2.21	13 2.25	3 —	7 2.49	— —	16 2.26	20 2.33	12 1.93	8 2.48	3 —
16	Psychoneuroses N Mean	17 3.38	19 4.04	26 4.01	11 4.18	37 3.85	43 3.76	30 4.09	16 3.67	14 4.48	13 4.31
17	Manic-depressive psy. N Mean	43 3.41	44 3.66	84 3.09	83 3.12	112 3.28	127 3.20	127 3.23	105 3.05	22 3.98	18 3.91
	171 Manic type N Mean	12 3.63	19 3.52	26 3.32	39 3.01	31 3.56	38 3.42	58 3.18	47 3.01	11 3.65	4 —
	172 Depressed type N Mean	17 3.27	18 4.01	32 3.03	31 3.24	46 3.14	49 3.11	49 3.32	40 3.10	9 4.31	8 3.77
	174 Mixed type N Mean	11 3.40	6 3.02	16 3.08	8 3.43	26 3.13	27 3.21	14 3.25	12 3.07	2 —	4 —

TABLE 20—(continued)

Code	Diagnosis	Male		Female		Total					Disch. betw. cens. (N = 211)
		Census I (N = 998)	Census II (N = 991)	Census I (N = 934)	Census II (N = 934)	Census I ^a ex. p.d. (N = 1,603)	Census I (N = 1,932)	Census II (N = 1,925)	Perm. pop. (N = 1,481)	Adm. betw. cens. (N = 444)	
18	Schizophrenia N Mean	498 2.98	528 3.00	470 2.80	498 2.83	896 2.94	968 2.89	1,026 2.92	889 2.82	137 3.00	74 3.82
	180 Unclassified N Mean	17 3.26	2 —	35 2.92	1 —	47 3.01	52 3.03	3 —	3 —	— —	2 —
	181 Simple type N Mean	96 2.88	102 2.91	59 2.97	63 2.77	144 2.93	155 2.91	165 2.86	140 2.77	25 3.33	17 3.91
	182 Hebephrenic type N Mean	124 2.87	111 2.85	133 2.38	134 2.45	239 2.65	257 2.02	245 2.63	236 2.60	9 3.60	4 —
	183 Catatonic type N Mean	39 2.53	48 2.44	43 2.63	44 2.84	73 2.68	82 2.58	92 2.63	79 2.50	13 3.34	11 3.41
	184 Paranoid type N Mean	217 3.15	250 3.16	190 3.03	234 3.01	378 3.14	407 3.09	484 3.09	418 3.00	66 3.67	36 4.02
	185 Other types N Mean	5 2.77	15 3.82	10 3.13	22 3.42	15 3.01	15 3.01	37 3.59	13 3.17	24 3.85	4 —
19	Paranoia and paranoid con- ditions N Mean	9 4.20	10 3.71	6 2.65	5 3.57	14 3.73	15 3.58	15 3.66	9 3.28	6 4.00	6 4.68

(Continued on page 22)

TABLE 20—(continued)

Code	Diagnosis	Male		Female		Total					
		Census I (N=998)	Census II (N=991)	Census I (N=934)	Census II (N=934)	Census I ^a ex. p.d. (N=1,003)	Census I (N=1,932)	Census II (N=1,925)	Perm. pop. (N=1,481)	Adm. betw. cens. (N=444)	Disch. betw. cens. (N=211)
21	Psy. with mental def. N Mean	48 2.80	50 2.88	66 2.75	51 2.89	91 2.90	114 2.77	101 2.89	87 2.88	14 2.96	10 3.53
23	Without mental disorder N Mean	74 3.29	73 3.32	39 3.24	27 2.80	97 3.39	113 3.27	100 3.18	73 2.95	27 3.73	17 4.29
	230 Unclassified N Mean	21 3.00	4 —	2 —	— —	19 3.12	23 3.18	4 —	4 —	— —	1 —
	232 Alcoholism N Mean	15 4.29	12 4.12	— —	— —	14 4.29	15 4.29	12 4.12	7 3.01	5 4.41	9 4.32
	234 Mental deficiency N Mean	23 2.91	35 2.92	31 3.24	23 2.69	45 3.14	54 3.10	58 2.83	49 2.80	9 3.08	2 —
	236 Psychopathic personal- ity N Mean	9 3.31	9 3.92	4 —	1 —	12 3.48	13 3.42	10 3.99	5 3.00	5 4.65	2 —

The age-corrected sample shows essentially similar characteristics.

Behavior Ratings in Relation to Diagnosis

This area presented a number of problems. The bias of the individual responsible for final diagnosis, the orientation of the institution, and the difficulty of sorting patients into Kraepelinian diagnostic categories are all responsible for some confusion in diagnosis. Older diagnoses may also be inaccurate, although most of the more obvious errors in diagnosis in this hospital have been changed. The major role of diagnosis in this institution is probably that of assistance in determining the treatment to be given a patient.

Table 20 is based upon the 1934 classification of mental disorder. Diagnoses in Census II have been fitted to this classification (4). Because of occasional marked differences between the male and female populations, characteristics of both are given. There are some differences in diagnoses between the two censuses. A gradual decline can be noted in the number of patients with a diagnosis of psychoses with syphilitic meningo-encephalitis (general paresis). Behavior ratings of the patients with this diagnosis have also become lower. More adequate treatment methods (both preventive and arrestive) may account for the declining numbers.

The number of patients diagnosed as psychoses due to alcohol has also shown a very marked decline. Behavior rating of patients with this diagnosis has become higher. Admissions to the new state hospital for alcoholics changed the admission rate at this hospital. Perhaps earlier and better treatment of alcoholics prevents a number of these psychoses from developing.

Under the diagnostic heading "psychoses with cerebral arteriosclerosis," approximately two-thirds of the patients are male. The mean behavior rating approximates the average of the hospital population. In the diagnostic group "senile psychoses," the number of females increased markedly between Census I and Census

II. This increase was especially noticeable in the diagnostic category "senile psychoses with simple deterioration." The very low average behavior rating of this diagnostic category suggests the possibility that the differential diagnosis between senile psychosis and psychosis with arteriosclerosis is made on the basis of age and lower behavioral level. (It should be noted that the death rate for the diagnosis "senile psychoses with simple deterioration" is much higher than that of the other groups.) The diagnosis of senile psychoses, paranoid type, has a behavior rating approximating the hospital average. Many more females are diagnosed as having involutional psychoses. The behavioral rating of patients with this diagnosis is considerably above that for the hospital population.

On Census II the senile and arteriosclerotic diagnostic categories show a definite increase in the absolute number of females. If females have a longer life span, and if the trend for increased senile admissions continues, one might anticipate that in the future there will be more female than male patients in the hospital.

Ages on Admission in Relation to Diagnoses

Mean ages of patients admitted between 1 June 1951 and 1 June 1952 in the categories "psychoses with cerebral arteriosclerosis," "senile psychoses," and "involutional psychoses" give further indication that these diagnostic differentiations may be made, at least partially, on the basis of age and behavior. The age differences between the three diagnostic categories are all significant. The mean age of male and female patients is approximately equal in each of the categories.

On the basis of some theories of the involutional period it would be anticipated that males diagnosed as involuntals would be about ten years older than females. Table 21 indicates that this is not the case. Further, males receive an involutional diagnosis infrequently. Apparently there is a strong male-female bias existing between the diagnostic categories "psychoses with

TABLE 21
AGES ON ADMISSION IN RELATION TO DIAGNOSES 1 JUNE 1951 THROUGH 31 MAY 1952

Code	Diagnosis	Male			Female		
		N ^a	Age	SD	N ^a	Age	SD
080	Psy. with cerebral arteriosc.	41	73.5	8.8	10	71.0	7.5
11	Senile psychoses	43	78.5	7.8	39	77.3	9.9
	III Simple deterioration	28	80.1	6.8	22	81.1	7.7
12	Involutional psychoses	20	53.4	6.5	41	51.6	6.6

^a The age was unknown for one male and one female patient.

cerebral arteriosclerosis" and "involutional psychoses." Extremely old patients are diagnosed as senile psychoses with simple dementia; younger male patients are diagnosed as psychoses with cerebral arteriosclerosis; and younger female patients are diagnosed as involutional psychoses. In view of gross differences in age and behavioral level in the categories, the possibility that diagnosis is not entirely on a physiological basis should be considered.

The diagnostic group "psychoneuroses" (Table 20) has a total of 43 patients. The mean behavior rating of these patients is considerably above that of the hospital as a whole. The most common subgroup diagnoses were anxiety-hysteria, reactive depression, and anxiety state. The psychoneurotics that are received in a state hospital may be much more maladjusted than those commonly encountered in private practice, since it is usually assumed that a neurotic is capable of adjusting within the community.

The group diagnosed as manic-depressive psychoses has a preponderance of female patients. Males are rated much higher than females, and the group rating is considerably above the hospital average. The behavior rating of those diagnosed as manic has shown a reliable decrease between Census I and Census II.

A majority of the patient population is included under the diagnosis "schizophrenia." On Census I, 968 patients had this diagnosis and 889 of these were in the hospital at the time of Census II. Only 74 patients with this diagnosis were discharged. This rate is below that expected on a statistical basis. Many more males than females received the diagnosis "simple schizophrenia." It will be noticed that the behavior rating of this group approximates that of the hospital. Of the 257 patients who had a diagnosis of schizophrenia, hebephrenic type, only four were discharged. Even though the admission rate is low, the discharge figure seems small for the relatively large number of these patients. (It is possible that this diagnosis was given more frequently in the past and current figures are not comparable.) The behavior rating of the hebephrenic patients was below that of the hospital. A diagnosis of schizophrenia, catatonic type, was given to 82 patients. The behavior rating for this group is also considerably below that of the hospital. On the other hand, the number of patients discharged ($N = 11$) from this group was fairly high considering the low number of admissions. The impression that patients diagnosed as catatonic schizophrenia have a fair prognosis agrees with views of others. The majority of patients diagnosed "schizophrenia" were of the paranoid type (407 patients). In this hospital this diagnosis is used as somewhat of a "catch-all." If a patient does not fit too well into another schizophrenic classification he is

usually diagnosed as a paranoid type. The average behavior rating of the paranoid schizophrenic is slightly higher than the hospital mean. There were a large number of patients in the undifferentiated classifications (e.g., such diagnoses as dementia praecox) and many of these were reclassified in the interval between censuses.

Many patients are diagnosed "psychoses with mental deficiency." The behavior rating of this group approximates the hospital average. The data indicate that a dual handicap of psychosis and mental deficiency does not prevent release of some patients to a noninstitutional environment.

In the category "without mental disorder," the patients diagnosed as mentally deficient form the largest subgroup. The behavior rating of this subgroup approximates the hospital mean. The alcoholics without mental disorder have the highest behavior rating in the hospital. A great many in this group were discharged. The psychopaths in this hospital also had a very high behavior rating. Approximately one-half of the individuals in this group were rediagnosed in the interval between censuses.

Most of the conclusions that can be reached from the diagnostic groups seem similar to those previously reported. If the data are adequate, some of the generally accepted concepts of regression (especially those regarding the diagnostic category "simple schizophrenia") need a reevaluation. Since hospitals differ in the diagnoses that are attached to various behavior patterns, one would expect differences in inter-hospital comparisons.

Behavior Ratings and Regression

An interesting speculative possibility concerns the relationship between the levels of behaviors rated on the behavioral census and those hypothesized in psychological theory. It has been widely held that latest learned behavior shows maximal regression and that earliest learned behavior shows minimal regression. Four psychologists and one psychiatrist from the University of Minnesota Hospital Psychiatric Unit were asked to rank independently the areas in the behavior rating scale in order, from earliest to latest learned behavior. (These clinicians did not know the results of the behavioral census.) The areas on the behavioral rating scale were also ranked from highest to lowest in terms of their mean values on Census I.

Agreements were computed by the rank-difference correlation method. The rankings of each clinician were compared with the rankings of each other clinician and also with the rankings determined from Census I. The ranks of each clinician were averaged for each individual area and a correlation was computed between these

TABLE 22

CORRELATIONS (RHO) BETWEEN FIVE CLINICIANS' RANKINGS OF BEHAVIOR AREAS AS TO EARLY VS. LATE LEARNING OF THE AREA; AND CORRELATION (RHO) BETWEEN CLINICIANS' RANKINGS AND AVERAGE BEHAVIOR RATING FOR EACH AREA IN CENSUS I

Clinician	Clinician					Average of 5 Clinicians
	I	II	III	IV	V	
I						
II	.69					
III	.63	.80				
IV	.58	.79	.76			
V	.61	.89	.90	.89		
Rank based on mean behavior rating of areas, Census I ^a	.67	.84	.68	.66	.69	.80

^a High behavior ratings were ranked low, and low behavior ratings were ranked high, so as to provide a measure of regression and thus yield positive correlations with the clinicians' rank-order of the areas with respect to early vs. late learning of the behavior area.

average ranks and the ranks from the behavioral census. The correlation by this method was .80. The ranking of the areas on Census I from high to low is as follows: K, Toilet Behavior; I, Psychomotor Activity; J, Speech; B, Response to Meals; H, Attention to Dress and Person; E, Response to Doctors, Social Workers, Psychologists; D, Response to Psychiatric Aides and Nurses; F, Response to Electric or Insulin Therapy; G, Occupational and Recreational Therapy; C, Response to Other Patients; A, Work. The ranking of the areas by the clinicians, with respect to early vs. late learning of the behavior area, is: B, Response to Meals; K, Toilet Behavior; I, Psychomotor Activity; J, Speech; H, Attention to Dress and Person; C, Response to Other Patients; E, Response to Doctors, Social Workers, Psychologists; D, Response to Psychiatric Aides and Nurses; A, Work; G, Occupational and Recreational Therapy; F, Response to Electric or Insulin Therapy. Although the

correlation is not high enough to be of predictive value, it indicates a definite relationship between psychological theory and the pattern of regression exhibited behaviorally.

Intercorrelation of Areas

The intercorrelations of the areas on the behavioral census were computed. With the exception of Area F, Attitude toward Insulin or Electric Therapy, the correlations of areas of the scale with the average (L) ranged from .69 to .87. The best predictor of the average rating was Area D, Response to Psychiatric Aides and Nurses. Area F showed relatively little correlation with other scales, the correlations ranging from .06 to .36. The other scales (excluding F) show intercorrelations of .42 to .78 with the most typical correlation being about .60. A factor analysis of this matrix found only one factor.

These correlations show a moderately high positive relationship to the general average and

TABLE 23

INTERCORRELATIONS OF AREAS ON THE L-M SCALE (FROM CENSUS I)

Area	A	B	C	D	E	F	G	H	I	J	K
A											
B	.66										
C	.67	.61									
D	.78	.69	.76								
E	.52	.51	.65	.63							
F	.26	.15	.32	.35	.36						
G	.63	.54	.67	.67	.51	.28					
H	.60	.72	.62	.66	.50	.13	.56				
I	.65	.65	.65	.68	.50	.25	.59	.65			
J	.54	.60	.64	.62	.56	.10	.52	.60	.62		
K	.54	.61	.44	.51	.43	.06	.42	.59	.55	.50	
L	.83	.81	.83	.87	.71	.34	.76	.81	.82	.78	.69

to each other. The fact that Response to Psychiatric Aides and Nurses (Area D) has the highest relationship to average behavior rating implies the operation of "halo" effect. Several factors should be considered in assessing the importance of this bias. The complexity of interpersonal relationships may mean that "less normal" patients are in fact lower in this area. "More normal" or "less normal" behavior may be a reflection of the patient's like or dislike for an aide (e.g., a feeding problem is more work for the aide).

The relatively modest intercorrelations of the areas (especially Area F) suggest that "halo" effect is probably minimal. In this connection it is often found, for example, that a patient will be a good worker and eat well, but will be poor at socializing with his peers. A wide variety of behavioral configurations can be observed.

SUMMARY

Two evaluations were made of the Fergus Falls State Hospital population, one in 1951 ($N = 1,932$), and one in 1953 ($N = 1,925$). (The 1951 survey is referred to as Census I, and the 1953 survey as Census II.) Data were gathered on the following items: Diagnosis, Present Age, Sex, Marital Status, Religious Affiliation, Physically Debilitating Conditions, Length of Hospitalization, Education, and Age at First Admission; and on the areas of the L-M Fergus Falls Behavior Rating Scale. Some characteristics of the normal population as given in the U. S. Census figures for the State of Minnesota for 1950 were used as a basis for comparison. These items were studied in an attempt to describe behavioral characteristics of a state hospital population. A description of this type has not previously been attempted on a state hospital population. It is hoped that results of this census will be of value to scientists, hospital administrators, and ultimately the lay public.

Conclusions reached from the study follow:

1. The hospital population is older than the normal population. The aged in the hospital population are increas-

ing. Hospital planning must take these facts into account.

2. Patients included in the census were most frequently admitted in early adult life (ages 20 to 40). However, a group of consecutive first admissions (from 1 June 1951 through 31 May 1952) had a mean admission age of 53.9 years. This figure agrees with a recent figure for *all* Minnesota state mental hospitals.

3. A majority of patients and a minority of normals are single. A correction for age greatly increases this difference.

4. There are more widowed people in the hospital than in the normal population. However, an age-corrected sample did not differ reliably from the normal population in this respect.

5. There probably is no difference between the hospital and normal populations with regard to religious preference.

6. There may be more physically debilitated people in the hospital than in the community. Advancing age may result in unfavorable personality changes as well as physical debilitation.

7. The mean length of hospitalization was 11.2 years. Therapeutic programming should take long-term patients into account. Females are hospitalized longer than males.

8. Even when an age-correction is made, hospital patients have less education than the normal population.

9. The hospital has a larger number of people diagnosed as feeble-minded (8.6%) than would be found in a normal population on the basis of intelligence test distributions.

10. The percentages of the various diagnoses in this hospital agree substantially with those found by others. A majority of the patient population is diagnosed as schizophrenic.

11. A comparison between diagnoses

of admissions (for a one-year period) and the hospital-census population indicates that the two groups differ markedly in some categories. The admissions are more frequently diagnosed as psychoses with circulatory disturbances, senile psychoses, involutional psychoses, psychoneuroses, manic-depressive psychoses, and alcoholism. The hospital-census populations are more frequently diagnosed as psychoses due to unknown or hereditary causes but associated with organic change, schizophrenia, psychoses with mental deficiency, and certain categories without mental disorder excluding alcoholism.

12. In regard to the general level of work performed in the hospital, males are rated as being reliably better than females. The average male is capable of holding a regularly assigned job under supervision. The average female seems to do a little work, with much urging, under constant supervision. Perhaps males are more accustomed to working in an industrial setting.

13. Males and females are approximately equal in their behavior at meals. The average patient is usually able to eat by himself, using utensils properly. Very few patients require special attention at meals.

14. Males are reliably better than females in social contacts with peers. Many males exhibit some spontaneity in making contacts and in initiating play or work of a fairly high order. The average female patient may be somewhat friendly and may even have an occasional friend.

15. On Census I males and females respond equally well to psychiatric aides and nurses. On Census II the average male has shown improvement. He has become reliably better than the average female and is much more apt to make

simple requests and to do things when asked.

16. Males are more responsive and friendly to non-nursing professional personnel than females.

17. Males are less antagonistic than females in the acceptance of the administration of electric or insulin therapy.

18. On Census I females tended to be better than males in response to occupational and recreational therapy. On Census II the male ratings show a statistically reliable improvement and become equal to the females. The average patient in occupational and recreational therapy will participate when asked and may occasionally show spontaneity.

19. Males show a tendency to be better than females in their attention to dress and person. The males improved at the time of Census II while the females were rated slightly lower (the difference between them became statistically reliable). The average patient has some interest in his appearance.

20. Males and females are approximately equal in level of psychomotor activity. Males showed a highly reliable increase in the level of their behavior on the second census. There are more female than male hyperactive wards in the hospital. The average patient seems to exhibit mainly purposeful behavior but has some erratic movements which are a result of the illness.

21. Males and females are approximately equal in speech, although females were somewhat higher on the first census. Between Census I and Census II the females showed a statistically reliable decrease in this area. This is the first reliable shift reported for females in this study. The average patient will speak in short, clear sentences if he has a request to make.

22. The males are reliably better than the females in their toilet behavior. This difference is of clinical significance. Both males and females improved on the second census. The total improvement was statistically reliable. The average behavior of the patient is normal except for being too neat or spending too much time at one thing and perhaps being rarely incontinent.

23. The males were reliably better than the females on both censuses on the average of all these areas of behavior. The authors feel that this difference is large enough to be of clinical importance. The males showed a reliable improvement between Census I and Census II. Females remained the same. Physically debilitated male and female patients have equal behavior averages.

24. Male admissions and discharges are higher in behavior rating than female admissions and discharges.

25. The consistent tendency for female patients to exhibit a lower level of behavior than male patients (and the results of a previous study in which females failed to hold their improvement on a *total-push* program) (1, 5) would indicate the need for more staff on the women's side of a state hospital. State hospital administrations might profitably reconsider the problems of the female patient with regard to such things as type of work, available clothing, and toilet space.

26. It is possible that an expanded therapeutic program was responsible for the improvement in behavior shown by the hospital population between Census I and Census II. Further censuses will be needed to substantiate this conclusion.

27. The average behavior rating of a group of student nurses ("normals") was approximately equal to that of the highest-level ward in the hospital. Some pa-

tients at this high behavioral level could leave the hospital if they were motivated to do so.

28. Patients who are physically debilitated have a lower behavior rating and are ten years older on the average than those who are not physically debilitated. The behavior rating of debilitated patients is fairly constant for all age groups.

29. Behavior of patients between the ages of 51 and 65 is slightly better than that of any other age group. Those patients over 86 show poorer behavior. Otherwise no definite relationship between age and behavior rating was discernible.

30. Up to age 65 patients show a trend toward higher behavior rating as age at first admission increases. Admissions after age 65 are poorer in terms of behavior. The physically debilitated population is ten years older than the *total* hospital population on age at first admission.

31. Up to twenty years of hospitalization, patients show a steady regression in behavior rating. Those who have been hospitalized between 26 and 45 years have a behavior rating higher than the average patient. Inclusion of physically debilitated patients would make the trend more pronounced. After 45 years of hospitalization there is a marked regression in behavioral rating, which may be due to aging. The hypothesis of continuing regression should be re-examined.

32. Type of religious affiliation (or lack thereof) does not seem to affect behavior rating.

33. Patients who have been divorced have the highest (most favorable) behavior rating; patients who have been married have the next highest; and those who are single tend to have the lowest rating. Single individuals may be more severely maladjusted than those who marry. As a group, physically debili-

tated patients are rated lower than the nondebilitated. Both groups have approximately the same relative ranking.

34. Behavior rating has a positive relationship to the number of years of school completed.

35. Because of the ambiguities that continually plague nosology, as well as the variations in biosocial environs, diagnoses present problems in intra- and inter-hospital comparison. More females than males receive manic-depressive diagnoses. The number of females has disproportionately increased between censuses in the categories of senile and involutional psychoses. If the trend toward older admissions continues, the longer life span of the female may eventually result in the presence of more female than male patients in the hospital. It is suggested that the differentiation between the categories of psychoses with cerebral arteriosclerosis, senile psychoses, and involutional psychoses is not entirely on a metabolic basis, but is, at least partially, a function of age, sex, and behavioral level.

36. The mean behavior ratings of simple schizophrenics approximate the average for the hospital. The average rating of hebephrenics was below the hospital mean. The catatonics' average was below the hospital mean but the rate of discharge was fairly high. Paranoid schizophrenics have an average rating slightly above the hospital mean.

37. Mentally deficient patients (with or without psychosis) are about average in behavior rating. Alcoholics, psychoneurotics, and psychopaths are very high in rated behavior.

38. A comparison of clinicians' rankings and the results of the behavioral census lends support to the hypothesis that regression in mental illness occurs from latest to earliest learned behavior.

39. The most typical intercorrelation of the ratings on the behavior areas was about .60 (all correlations were positive). Although behaviors tend to be positively interrelated, this finding supports the observation that patients may function at different levels in various areas. Factor analysis yielded only one factor.

APPENDIX

L-M FERGUS FALLS BEHAVIOR RATING SHEET (2, 3)

Name	Hospital Number	Age	Sex	Marital Status	Date
Diagnosis	Time on Present Ward	Ward	Religious Preference		

A. Work.

- (1)⁶ Does no work—refuses—extremely negativistic.
- (2) Does a little work with a lot of urging. Constant supervision is necessary.
- (3) May have a regularly assigned job—and supervision may be necessary.
- (4) Enthusiastic participation in all types of work—asks for work.
- (5) Normal interest in work—i.e., interested in some kinds of work more than others (will do other kinds than main interest if called upon to do so).

(If works, at what, and how many hours a day?)

B. Response to Meals.

- Has to have special attention, as eats too much, or is spoon-fed or tube-fed.
- Eats by self, is sloppy—may need coaxing.
- Eats by self using knife, fork, and spoon properly—may show some finickyness.
- Passes and asks for things to be passed, but will not carry on table conversation.
- Would not stand out among normal people for eating habits.

C. Response to Other Patients.

- Stays all alone or may strike out at other patients.
- Will be with other patients only for a short while and with urging.
- Some signs of friendliness—speaks to patients—may have a friend.
- Some spontaneity in making contacts with other patients. May initiate play or work of a social, relatively high type, e.g., card games, washing dishes, etc.
- Helpfulness expressed toward other patients—or non-hostile recognition of their being mentally ill and making allowances.

(If stays alone or strikes, which one?)

D. Response to Psychiatric Aides and Nurses.

- Negativistic—hostile (can include strik-

⁶The number preceding each statement does not appear on the scale. Only a blank space is provided for a check mark. A clerk assigns the numerical value.

- ing)—doesn't do anything requested.
- Will do a few things if asked or pushed—shows no open hostility.
- Will do most things when asked—will ask for simple things—"I want my toothbrush."
- Extremely cooperative—will do anything when asked.
- Normal give-and-take relationship. Speaks spontaneously to nurses about things of not immediate importance, e.g., weather, baseball games, etc.

E. Response to Doctors, Social Workers, Psychologists.

- Hostile.
- Passively negativistic (would rather not have anything to do with them but will not resist).
- Will speak when spoken to.
- Seeks advice.
- Understands, accepts, and asks for therapy.

F. Response to Electric or Insulin Therapy.

- Hostile.
- Anxious, apprehensive, but not overly hostile.
- Passively accepts.
- Accepts positively—(may say, "I feel better after").
- Asks for, understands necessity for.

G. Occupational Therapy and Recreational Therapy (walks don't count).

- Does not participate at all—negativistic—hostile.
- Participates with urging for short periods.
- Participates when asked—some spontaneity.
- Shows interest—participates in all types wholeheartedly without discriminating very much between different types—looks forward to.
- Interested in many varied activities—normal selectivity (likes some kinds more than others).

(What is patient most interested in?)

H. Attention to Dress and Person.

- Has to be dressed—needs special attention of one kind or another.
- Dresses self but is sloppy.
- Some interest in looks—too much lipstick, fairly neat.
- Cares about looks and dress—will ask for make-up or shaving equipment inconsistently (not an over-all balance).
- Normal (for culture)—would not stand out in a crowd.

I. Psychomotor Activity (not including going to the bathroom or meals).

- Stays in one place unless pushed—or hyperactive (e.g., seclusion necessary).
- Moves around a little (one chair to another) or, if hyperactive, the activity is not of a type making seclusion or other restrictions necessary.
- Some activity resulting from the influence of the illness (moves around because voices say to) and some purposeful behavior.
- Behavior mainly purposeful. Still moves around a little fast or a little slow.
- Normal activity—would not stand out among normal people.

(If hyperactive or stays in one place, which one?)

J. Speech.

- Mute or speaks a lot but doesn't make sense.
- A few words that make sense ("yes" or "no").

- Speaks in short, clear sentences—"Can I have my toothbrush."
- Speaks normally except a little fast or slow.
- Speaks normally.

(If mute or senseless talk, which one?)

K. Toilet Behavior.

- Untidy anytime during the day and/or more than twice a week nightly.
- Untidy once or twice a week nightly—brushes teeth and washes only when told to do so.
- Not untidy—toilet behavior somewhat sloppy—brushes teeth and washes once a day without being told.
- Toilet behavior normal except for being too neat, or spending too much time at one thing, or occasionally sloppy.
- Toilet behavior normal.

List below anything physically wrong with patient's arms or legs.

List below anything physically wrong with patient's hearing, sight, and speech organs.

List below any physical illness that patient has.

REFERENCES

1. FJELD, S. P., LUCERO, R. J., & RECHTSCHAFFEN, A. Cross-validation and follow-up of a state hospital total push program for regressed schizophrenics. *J. clin. Psych.*, 1953, 9, 394-395.
2. LUCERO, R. J., & MEYER, B. T. A behavior rating scale suitable for use in mental hospitals. *J. clin. Psych.*, 1951, 7, 250-254.
3. MEYER, B. T., & LUCERO, R. J. A validation study of the L-M Fergus Falls Behavior Rating Scale. *J. clin. Psych.*, 1953, 9, 192-195.
4. NATIONAL COMMITTEE FOR MENTAL HYGIENE. *Statistical manual for the use of hospitals for mental disease*, Tenth Edition. 1790 Broadway, New York 19: 1942.
5. SINES, J. O., LUCERO, R. J., & KAMMAN, G. R. A state hospital total push program for regressed schizophrenics. *J. clin. Psych.*, 1952, 8, 180-193.
6. U. S. BUREAU OF THE CENSUS. *Religious bodies*, 1936. 3 vols. Washington, D. C.: U. S. Government Printing Office, 1941.
7. U. S. BUREAU OF THE CENSUS. *U. S. Census of population: 1950*. Vol. II, *Characteristics of the population*, Part 23, Minnesota, Chapters B and C. Washington, D.C.: U. S. Government Printing Office, 1952.

(Accepted for publication March 2, 1957)

Psychological Monographs: General and Applied

The Retinal Size of a Familiar Object as a Determiner of Apparent Distance¹

Walter C. Gogel, Bryce O. Hartman and George S. Harker
Army Medical Research Laboratory, Fort Knox, Kentucky²

I. INTRODUCTION

THE present study is concerned with the retinal size of a familiar object as a cue to distance. A distinction is sometimes made between apparent absolute distance and apparent relative distance. Apparent absolute distance is the apparent distance of an object from the observer. Apparent relative distance is the apparent depth between objects without reference to the position of the observer. The characteristic which discriminates between these two types of apparent distance is the reference point or point of zero distance with respect to which the perception occurs. If this reference point involves the apparent position of the observer, an absolute-distance perception has occurred. If this reference point is independent of the apparent position of the observer and only involves objects external to the observer, relative-distance perception is involved.

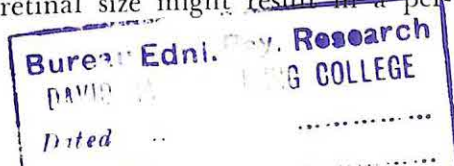
One of the factors which is considered to be a cue to the perception of absolute distance is the retinal size of familiar objects. For a constant physical size, the retinal (angular) size of an object decreases with increasing distance from the observer. Familiarity with a certain physical size of an object at various distances might result in a distance being perceptually ascribed

to each retinal size of the familiar object. When the observer is presented with a familiar visual object of a particular retinal extent, the perception might be that the object is at a particular distance, with this perceived distance being related to the retinal size of the object (5, 6). This will be called the *size-distance hypothesis*. According to this viewpoint, a familiar visual stimulus of a particular retinal size will be seen at the distance position which a normalized object of the same familiar category would have to occupy in order to have the particular retinal size. Within limits, if the retinal size doubles, the perceived distance should halve; if the retinal size is halved, the perceived distance should double, etc. It is a problem of this study to investigate whether the retinal subtense of a familiar object can act as a determiner of the apparent absolute distance of that object from the observer.

If two similar familiar objects of different retinal size are presented simultaneously, the one with the smaller retinal subtense may appear to be more distant than the other. This might occur even though neither object is perceived as being at any particular absolute distance. In this case, the perceived-distance difference of the two objects would involve a relative- not an absolute-distance perception. If the two objects were presented successively, the same results might occur and the same interpretation be applied to these results. In the case of successive presentation, the memory of the retinal extent of the first object could be compared with that of the second object. With either the simultaneous or successive presentation of objects, relative retinal size might result in a per-

¹ The authors wish to thank Kay Inaba, Robert E. Page, and John J. Cox for their help in collecting and analyzing the data.

² The opinions or conclusions contained in the present report are those of the authors. They are not to be construed as reflecting the views or endorsement of the Department of the Army.



ceived relative distance. *In investigating whether the retinal subtense of a familiar object is an effective cue to its absolute distance, the relative-size cue must not be available between objects presented in the test situation.* This indicates two requirements for determining the relationship between familiar retinal subtense and perceived absolute distance: (a) a test situation must be used in which only *one object* is presented visually to the subject; and (b) the different retinal subtenses of the familiar object must be presented to different subjects.

The problem of determining whether an absolute- or a relative-distance perception (or both) are involved in a particular distance perception is difficult. Pertinent to this problem is a consideration of the response by which the distance perception is measured. The first requirement, stated at the close of the preceding paragraph, for finding the relation between the retinal subtense of a familiar object and perceived absolute distance suggests that a nonvisual method of measuring perceived absolute distance be used. In the present study a motor performance was employed. The subject was asked to throw darts to the distance of an object without being able to see the results of the throw. This task would ordinarily require that the subject judge the distance of the object from himself. However, even the results of this task must be interpreted with caution. If, for example, an object appears indefinitely located in distance, a subject, when given the task of throwing to the distance of this object, may make some throwing response. If, subsequently, an object is presented which appears to be less distant than the first object, the subject may throw to a lesser distance. This may occur even though both the first and the second object are perceptually in-

definitely located in their distance from the subject. In this case, the difference between the throwing to the two successively presented objects would indicate the perceptual depth between them. The change in the throwing scores could be considered a measure of relative and not of absolute distance perception. Consequently, for any subject, only the throwing to the object which is *first* presented may be considered in measuring the perceived absolute distance of that object. This is in agreement with the second requirement stated previously.

The experimental evidence for the occurrence of relative- and absolute-distance perception as a function of retinal size has been reviewed by Ittelson (3). Several studies have seemed to indicate that the retinal size of a familiar object is a cue to its absolute distance. Hastorf (1) had subjects adjust the size of a monocularly observed object until it appeared to be equidistant with a post in a binocularly observed field. He found that the adjusted size of the monocular object differed directly as a function of its assumed physical size, in order for it to appear at the same distance position as the post. He also found that the apparent distance position of the monocular object with respect to the post changed as its assumed size was changed. Ittelson and Ames (4) were interested in determining the effect of apparent distance upon accommodation and convergence. The apparent distance of a playing card was measured by adjusting a movable post to the perceived distance of the card. The retinal size of the playing card was varied by the experimenters. It was found that, for the same physical distance of the playing card, the adjusted position of the post varied inversely with the retinal subtense of the playing card. Ittelson (3) had subjects adjust a movable package of cigarettes or a checkerboard design to the apparent distance position previously occupied by an experimental object. In one experiment two sizes of playing cards, ink blots, and diamond-shaped figures were used as experimental objects. The assumed size and physical distance of the experimental object was kept constant while its retinal size was varied. In a second experiment using playing cards, a typewritten letter, and a match box as experimental objects, either or both assumed size and retinal size were varied with physical distance held constant. In both experiments, when familiar experimental objects were employed, the movable object was adjusted,

in general, to a distance position which a normal-sized object of the same kind as the experimental object would have to occupy, in order to have the same retinal subtense as the experimental object.

The results of these three studies seem to indicate that the retinal size of a familiar object can determine its apparent absolute distance. But, in none of these studies has the possibility been eliminated that the perception which was measured was a perception of relative, not absolute, distance. The method of measuring perceived absolute distance involved an apparent distance judgment between an experimental object and another object (a comparison object) with both objects visually presented. In the experimental field the only cue to distance was the retinal size of the experimental object. *The comparison object was presented in a generally well-structured comparison field. Regardless of whether the experimental and comparison objects were presented simultaneously or successively, this method permitted relative-size judgments to be used in making the distance judgment.*

Consider the case in which a familiar object was present in both the experimental and comparison field. For example, in one part of the study by Ittelson (3), a package of cigarettes in the comparison field was adjusted to the apparent distance position previously occupied by a $1\frac{1}{2}$ -size playing card in the experimental field. The resulting distance adjustment of the package of cigarettes was to a distance which was approximately two-thirds of the physical distance of the playing card (3, p. 63). This result can be explained by assuming that the subject adjusted the distance of the cigarettes until a ratio of the retinal size of the cigarette package and playing card was reached which was that normally experienced by the subject when ordinary objects of these types were at the same distance from himself. If this explanation is correct, the perception of absolute distance was not involved in the adjustment.

Consider the case in which a nonfamiliar object was used in the comparison field with a familiar object in the experimental field. For example, in the study by Ittelson and Ames (4) the comparison object was a movable post and the experimental object was a playing card. The study by Hastorf (1) also is an example of this case. The assumption of the observer that the experimental object was, for example, a ping-pong ball made the experimental object a familiar object. To the extent that the comparison field was sufficiently well-structured to provide information as to the physical size of the post, a relative size judgment between the post and the familiar experimental object could be used in adjusting the post and experimental object to apparent equidistance.

These three studies do not meet the first requirement given previously for testing the relation between size and perceived absolute distance. The second requirement was also not met, in that the same subjects were used in more than one test situation. Subject comparisons of the retinal sizes of experimental objects between subsequent test situations may be expected to result in the perception of relative distance. That this will occur even when the experimental objects are not familiar is suggested by the results which Ittelson obtained when ink blots and diamond-shaped figures were used as experimental objects. The apparent distance of the experimental object, as determined by the apparent equidistance adjustment of the comparison object, was less for the larger than for the smaller experimental object of similar shape (3, p. 63). The adjustment of the comparison object in the initial situation would be to an indifferent distance. The adjustments of the comparison object in subsequent situations, involving an experimental object of the same shape but different size, would be inversely related to the size change of the experimental object.

II. APPARATUS AND PROCEDURE

In the present study, the throwing of darts was used to measure the perceived distance of an object in situations in which, ideally, *only the retinal size of a familiar object was present as a possible cue to distance*. From the experimental result of thrown distance, perceived distance was to be inferred. For this reason, it was considered desirable to specify the magnitude of the throwing response as a function of the perceived distance. This was experimentally determined by using situations in which a variety of cues to distance were available. These situations are called "full-cue situations." The situations in which it was attempted to eliminate all cues to depth but that of retinal subtense are called "reduced-cue situations." It is assumed that in the full-cue situations the distance of the object from the observer was correctly perceived. Any difference in the full-cue situations between the physical and the thrown distance was attributed to a lack of one-to-one correspondence between visually per-

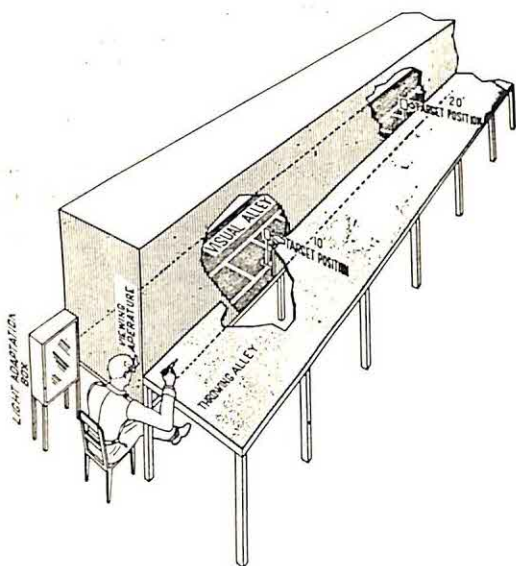


FIG. 1. Schematic representation of the throwing and visual alley for the full-cue situations.

ceived distance and thrown distance, and not to any inaccuracy in the perceived distance. The relation between thrown and visually perceived distance, which was experimentally determined from the full-cue situations, was used to convert thrown distance to perceived distance in the reduced-cue situations.

A. The Reduced-Cue Situations

1. Apparatus

The experimental situations of this study were presented in a visual alley separated from an adjacent throwing alley by a partition of dark cloth. Figure 1, which illustrates the full-cue situations, can also be used to discuss the reduced-cue situations. Both the visual and throwing alley were 4 feet wide and 40 feet long. A small rectangular aperture (1 inch by 5 inches), located 15 inches above the surface of the visual alley and at the same height as the surface of the throwing alley, allowed the subject to see into the visual alley. The surfaces of the visual alley and the surface containing the viewing aperture were dark. The ladder pattern shown on the floor of the

visual alley in Fig. 1 was present only in the full-cue situations.

The familiar object used in the reduced-cue situations was produced by seven-of-spades photographic transparent positives. One transparency was of a normal-sized card ($2\frac{1}{4}$ inches by $3\frac{1}{2}$ inches) and the other was of a double-sized card ($4\frac{1}{2}$ inches by 7 inches). Each of the two transparencies was fitted on the face of a light box and transilluminated with white light. No light from the box was visible except that passing through the transparency. When viewed in an otherwise dark room, a lighted transparency appeared to be an illuminated playing card and will be referred to as a playing card throughout this study. When a playing card was presented in the visual alley, the bottom of the card was at the height of the floor of the throwing alley.

Only one card was present in any one situation. Four reduced-cue situations were used. These will be designated by letter as follows:

- A—A double-sized card at 10 feet.
- B—A normal-sized card at 10 feet.
- C—A double-sized card at 20 feet.
- D—A normal-sized card at 20 feet.

Considerable care was taken with the reduced-cue situations so that no part of the visual alley or any object except the single playing card was visible to the subject when he looked through the viewing aperture. Black baffles were placed in the visual alley so that light from the playing card which fell on the floor and walls of the alley would not be visible. As an additional precaution, the subject was light-adapted before looking through the viewing aperture. The light-adapting surface (see Fig. 1) had a brightness of 11 foot-lamberts. The brightness of each playing card was 1.1 foot-lamberts.

The cloth partition between the visual and throwing alley prevented the subjects from seeing where the thrown darts landed. This was necessary since, if the thrown darts had landed in the field of view, relative-distance judgments between the card and dart could have occurred. The darts used in the throwing were 6 inches long and weighed approximately 11 grams. The distance to which the darts were thrown in the throwing alley was measured to the nearest centimeter.

2. Procedure

Eighty men who were right-handed and had at least 20/20 corrected vision in their right eyes were used as subjects. Prior to being subjects, these men were not informed of the purpose of the experiment. In the reduced-cue situations the viewing was monocular with the left eye of the subject covered. Before looking into the visual alley, the subject was given 20 practice throws to accustom himself to the throwing task. The throwing alley was lighted for this purpose. Following this, all lights except that from the adaptation surface and playing card were turned off. A shutter covered the viewing aperture. A mark on this shutter allowed the subject to orient his head so that his eye would be in the correct viewing position when the shutter was raised. The subject viewed the adaptation surface for approximately five seconds, after which he oriented his head with respect to the shutter mark and the adaptation surface was turned off. The shutter was raised and the subject looked at the playing card. By questioning, it was ascertained that the subject saw the stimulus as a seven-of-spades playing card. The subject was reminded of his task which was to throw a dart so that it landed at the distance of the playing card. He was also reminded that the height of the bottom of the card was the same as that of the floor of the throwing alley. The subject was handed a dart, which he threw down the throwing alley. The shutter was then closed, the adaptation light turned on, and the subject again light-adapted. This procedure of light-adaptation, looking into the viewing alley, and throwing a dart was repeated until 10 successful (scorable) throws were obtained for the particular subject in the particular situation. Occasionally, a dart either did not land on or stick into the floor of the throwing alley. Such a throw was not considered to be successful and was not scored.

Each of the 80 subjects was presented with each of the four situations, A, B, C, and D. A four-minute interval occurred between the end of the presentation of one situation and the beginning of another. The order in which the four situations were presented was varied between subjects. Since 24 orders were possible, these 24 orders were used three times (totaling 72 subjects), with four orders and their inverse selected again from the 24 possible orders, such that each situation (A, B, C, or D) was presented first equally often. Thus, of the 80 subjects, 20 were first presented with the reduced-cue situation A, 20 other subjects were first presented with B, 20 others with C, and 20 others with D.

Following the 10 scorable throws to the apparent distance of the card in a reduced-cue situation, the subject was asked to estimate the distance to the card in feet.

B. The Full-Cue Situations

1. Apparatus

For the full-cue situations, the fluorescent lights above the visual alley were turned on. There were no lights used in the throwing alley and the results of the dart throwing were not visible to the subject. The walls and floor of the visual alley were visible when the subject looked through the viewing aperture. A 38-inch-wide strip of brown paper was placed on the floor of the visual alley. Fastened to this strip of paper was a ladder pattern formed from a 1-inch-wide white tape. The sides of this pattern were separated by 28 inches, with the horizontal extensions occurring at 18-inch intervals. This pattern, part of which can be seen in Fig. 1, extended from the subject to a distance of 26.5 feet. A normal-sized seven-of-spades playing card (not a transparency) was mounted on a stand and placed at either 5, 10, 15, 20, or 25 feet from the subject. The normal-sized card located at each of these five distances constituted five full-cue situations. The bottom of the playing card was always at the height of the floor of the throwing alley. The subject binocularly observed the visual alley, in-

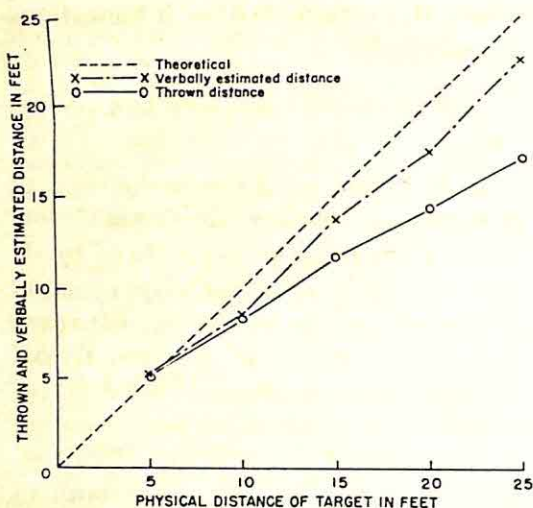


FIG. 2. Relation between physical and behavioral distance from all presentations of the full-cue situations.

cluding the playing card, through the viewing aperture. The light adaptation surface was not used in the full-cue situations.

2. Procedure

The same 80 subjects were used in the full-cue situations as had been used previously in the reduced-cue situations. The full-cue situations were always presented after the reduced-cue situations so that the subjects would be naive when they were first presented with the reduced-cue situations. Each subject completed 10 scorable dart throws to the apparent distance of the card in each of the full-cue situations. As in the reduced-cue situations, the task of the subject was carefully explained and he was reminded that the bottom of the card was at the height of the floor of the throwing alley. The shutter was closed over the viewing aperture between throws. A four-minute interval occurred between the end of one full-cue situation and the beginning of another. All five full-cue situations were presented to each subject, but the order in which the five full-cue situations were presented was different for each subject. Forty orders and their inverse were used. There were 16, 16, 17, 15, and 16 different subjects who were first presented with the target in the full-cue situations at 5, 10, 15, 20, and 25 feet respectively. For each subject, the time interval between the end of the last reduced-cue situation and the beginning of the first full-cue situation was 7 minutes. The experimental order for each subject throughout the entire experi-

ment was specified by randomly pairing the order in which he was to see the reduced-cue situations with one of the 80 orders of presenting the full-cue situations.

Following 10 scorable throws to the apparent distance of the card in any full-cue situation, the subject was asked to estimate the distance to the card in feet.

III. RESULTS

A. The Full-Cue Situations

The results from the full-cue situations are presented first, since these results produced the calibration equations used in the analysis of the reduced-cue situations. The average throwing scores and verbal estimates from all presentations of the full-cue situations as a function of the physical distance of the playing card are shown in Fig. 2. The dotted line represents the results which would have been obtained if the throwing and verbal scores were the same as the physical distances of the playing card. Each thrown distance represented by a point on the ordinate of Fig. 2 is a mean of 80 scores, one from each subject, where each score is an average of 10 dart throws. Each verbal estimate represented by a point on the ordinate is a mean of 80 verbal estimates with one estimate from each of the 80 subjects. The graph involving the verbal estimates is closer to the dotted line than is the graph involving the thrown distance scores. However, the average of the between-subject standard deviations of the verbal estimates (12.9 feet) is larger than that from the throwing scores (2.1 feet). The average standard deviation of 12.9 feet is inflated by the verbal responses of one subject. When the verbal data of this subject are excluded, this standard deviation reduces to 4.2 feet. In order to estimate the precision with which the throwing method differentiated between target positions, standard errors of differences were computed for the differ-

ences between throwing scores when the targets were physically separated by five-foot intervals. Averaging the four resulting standard errors gave a value of .2 feet. This suggests that, in the full-cue situations, a distance difference of .3 feet could have been statistically differentiated by the dart-throwing at the .05 level of confidence (using the single-tailed t test).

B. *The Reduced-Cue Situations*

As was mentioned previously, the relation between thrown and visually perceived distance was determined from the results of the full-cue situations and was used to convert thrown distance to perceived distance in the reduced-cue situations. This was accomplished as follows: By the method of least squares a linear relation was calculated, for each subject, between the dart-throwing scores and the physical distance of the card in the full-cue situations. It was assumed that physical and perceived distance was equivalent in the full-cue situations. This resulted in a linear equation for each subject which expressed throwing distance as a function of perceived distance. This function is called the calibration equation for a subject. Each throwing score (a mean of 10 throws) from each subject in each reduced-cue situation was transformed by the calibration equation to a "perceived-distance score." Actually, the perceived-distance scores can be considered to represent a perceived absolute distance only if the perception of absolute distance occurred in the reduced-cue situations. If the perception of an absolute distance was absent in the reduced-cue situations, the throwing would be to an indifferent distance. However, in order to investigate whether the perception of absolute distance varied with changes in the retinal size of a familiar

object, this transformation is logically to be preferred.

1. *First-Presented Reduced-Cue Situations*

As discussed in the introduction, in determining whether the retinal subtense of a familiar object is a cue to absolute distance, only one object should be visually present in the test situation, with similar objects of different retinal subtense being presented to different subjects. The results from each group of 20 subjects, who saw a particular reduced-cue situation first, meet both of these requirements. Therefore, the results from these first presentations of the reduced-cue situations were analyzed separately from the remaining results from the reduced-cue situations.

According to the size-distance hypothesis, the playing card in the reduced-cue situations A, B, C, and D should appear at 5, 10, 10, and 20 feet respectively. Let P_{A_1} , P_{B_1} , P_{C_1} , and P_{D_1} represent the perceived distance of the playing card in the first presentations of the reduced-cue situations A, B, C, and D. The subscript "1" refers to first presentations. The predictions from the size-distance hypothesis are that $P_{A_1} = 5$ feet, $P_{B_1} = 10$ feet, $P_{C_1} = 10$ feet, and $P_{D_1} = 20$ feet. A more general but less stringent hypothesis relating retinal size and perceived absolute distance is that $P_{A_1} < P_{B_1} = P_{C_1} < P_{D_1}$. If the results from this study do not support the latter hypothesis, they also do not support the former hypothesis. The analysis of the results will be mainly with respect to the latter, more general, hypothesis.

Table 1 gives the average results which were obtained from the first presentations of the reduced-cue situations. The entries in the column labeled "Thrown-Distance Score" are based upon

TABLE 1
RESULTS IN FEET FROM THE FIRST PRESENTATIONS OF THE REDUCED-CUE SITUATIONS
($N = 20$)

Reduced-Cue Situation	Thrown-Distance Score		Perceived-Distance Score		Verbal Report		Size-Distance Prediction	Physical Distance
	Mean	SD	Mean	SD	Mean	SD		
A	16.4	4.4	22.4	7.1	22.4	18.0	5	10
B	14.8	4.6	22.1	9.1	14.3	8.1	10	10
C	17.7	5.0	25.9	10.1	19.0	12.7	10	20
D	19.2	5.1	29.9	9.9	20.2	13.9	20	20

20 scores, with one score from each of 20 subjects, where each score is a mean of 10 throws of the darts. The entries in the column labeled "Perceived-Distance Score" were obtained by applying the appropriate calibration equations to the thrown-distance scores. The column labeled "Size-Distance Prediction" gives the distances at which the playing cards should have been seen according to the size-distance hypothesis. The column labeled "Physical Distance" gives the actual distances at which the playing cards were placed.

From Table 1 it appears that neither the thrown distances, the perceived distances, nor the verbally reported distances are in agreement with all the relationships involved in $P_{A_1} < P_{B_1} = P_{C_1} < P_{D_1}$. The statistical significance of the differences between the perceived-distance scores of Table 1 was determined by a simple analysis of variance.

TABLE 2
SUMMARY OF ANALYSIS OF VARIANCE OF THE PERCEIVED-DISTANCE SCORES FROM THE FIRST PRESENTATIONS OF THE REDUCED-CUE SITUATIONS

Source of Variance	df	Mean Square	F
Between Situations	3	248,968	3.22
Within Cells	76	77,346	
Total	79		

The results of this analysis are shown in Table 2. The obtained F is significant between the .05 and .01 level of confidence.

The significance of the differences between pairs of perceived distances P_{A_1} , P_{B_1} , P_{C_1} , and P_{D_1} were tested using the within-cells variance from the analysis of variance in calculating the error term. The results of these tests for single-tailed t probabilities are shown in Table 3. The direction of the subtractions in the column labeled "Test" was derived from

TABLE 3
SIGNIFICANCE OF THE DIFFERENCES BETWEEN PERCEIVED-DISTANCE SCORES FROM THE FIRST PRESENTATIONS OF THE REDUCED-CUE SITUATIONS

Test	Qualitative Differences	Difference in Feet	t	p^a $df = 1/76$
1. $P_{B_1} - P_{A_1} > 0$	Angle	- .3	.10	Reversal
2. $P_{D_1} - P_{C_1} > 0$	Angle			
3. $P_{C_1} - P_{B_1} > 0$	Distance	4.0	1.38	—
4. $P_{C_1} - P_{A_1} > 0$	Angle and Distance	3.8	1.32	—
5. $P_{D_1} - P_{B_1} > 0$	Angle and Distance	3.5	1.23	—
6. $P_D - P_{A_1} > 0$	Angle and Distance	7.8	2.71	< .01
7. $(P_{D_1} - P_{A_1}) - (P_{C_1} - P_{B_1}) > 0$	Angle Corrected for Distance	7.5	2.61	< .01
		3.7	.91	—

^a Single-tailed probability of occurrence.

the predictions of the size-distance hypothesis. The word "Reversal" in the column labeled "*p*" means that the obtained difference was opposite in direction to that predicted from this hypothesis. The column labeled "Qualitative Differences" in Table 3 indicates whether the playing cards in the two reduced-cue situations differed from each other in either or both angular (retinal) size and distance from the subject. It will be observed from test 1 and test 2 of Table 3 that, by itself, a decrease in the visual angle of the playing card did not produce a statistically significant increase in perceived distance. But, in 2 out of 3 cases, a decrease in visual angle accompanied by an increase in physical distance did produce a significant increase in perceived distance (see tests 4, 5, and 6). It is reasonable to expect that, in spite of the attempt to eliminate all distance cues except that of visual angle, some residual effect of other distance cues remained in the reduced-cue situations. This effect would be reflected in the results of test 3 in which the visual angle of the playing card was held constant but its physical distance was changed. As shown by the results of test 3, when the physical distance of the playing card was increased the perceived-distance score increased, but not by a statistically significant amount. Before concluding that the significant differences found in tests 5 and 6 can be attributed to the change in visual angle of the playing card, it must be demonstrated that these results cannot be attributed to the possible presence of other distance cues. For this reason, the difference found in test 3 was subtracted from the differences found in tests 5 and 6. Actually, the subtraction of the difference found in test 3 from that found in test 5 is unnecessary, since the resulting

difference is the same as that from test 2. The subtraction of test 3 from test 6 is shown as test 7 in Table 3. It will be seen that the difference in test 7 was not significant at the .05 level of confidence. It can be concluded that the perceived-distance scores of Table 2 provide no evidence that the retinal subtense of the playing card was used as a cue to determine the perception of the absolute distance of the card.

An analysis of variance also was performed on the throwing scores of Table 1. The *F* value of this analysis was significant beyond the .05 level of confidence ($F = 3.07$). The significance of the differences between pairs of these throwing scores was tested by the procedure indicated in Table 3. The single-tailed probabilities resulting from the *t* tests were similar to those obtained from the analysis of the perceived-distance scores, with the only difference of any importance being that test 3 was significant beyond the .05 level of confidence ($t = 1.93$) when the throwing rather than the perceived-distance scores were used. Neither the throwing scores nor the perceived-distance scores offer support for the hypothesis that the retinal size of the familiar object was used as a cue to its absolute distance.

2. All the Reduced-Cue Situations

It will be recalled that the reduced-cue situations A, B, C, and D were each presented successively to each of the subjects. So far, the analysis of the data has been only with respect to the results from the first presentations of the reduced-cue situations. The average results from all presentations of each of the reduced-cue situations are given in Table 4. Each mean and standard deviation of Table 4 is based upon 80 scores, with one score from each of the same 80 subjects. The

Bureau Ednl. Resear.

DAVID M. ALING COLLEGE

Dated

Page No.

TABLE 4
RESULTS IN FEET FROM ALL PRESENTATIONS OF THE REDUCED-CUE SITUATIONS
($N=80$)

Reduced-Cue Situation	Thrown-Distance Score		Perceived-Distance Score		Verbal Report		Size-Distance Prediction	Physical Distance
	Mean	SD	Mean	SD	Mean	SD		
A	12.1	3.8	16.5	8.6	12.4	8.0	5	10
B	14.9	1.6	21.3	7.9	16.3	11.9	10	10
C	15.6	1.7	22.8	9.5	21.1	13.0	10	20
D	19.0	1.5	28.4	8.0	26.1	19.7	20	20

calibration equations were used to convert throwing scores to perceptual scores. An analysis of variance was performed on these perceived-distance scores with the results shown in Table 5. The F obtained for the main effect (situations A, B, C, and D) was significant beyond the .001 level of confidence. The results of the one-tailed t test of the significance of the differences between pairs of the perceived-distance scores of Table 5 are given in Table 6. It will be observed that all the differences in Table 6 are significant beyond at least the .025 level of confidence. The perceived-distance scores varied inversely with the retinal size of the playing card between successive presentations of the reduced-cue situations. As indicated by tests 1, 2, and 7 of Table 6, these results cannot be accounted for by the effect of distance cues other than changes in retinal size.

A parallel analysis was completed for

TABLE 5
SUMMARY OF ANALYSIS OF VARIANCE OF THE PERCEIVED-DISTANCE SCORES FROM ALL PRESENTATIONS OF THE REDUCED-CUE SITUATIONS

Source of Variance	df	Mean Square	F
Between Situations	3	1,775,177	81.47
Between Subjects	79	215,149	9.87
Situations \times Subjects	237	21,789	
Total	319		

the throwing scores of Table 4. The resulting F and t values were very similar to those which occurred with the perceived-distance scores in Tables 5 and 6. Both the throwing scores and the perceived-distance scores support the hypothesis that the visual angle of the playing card significantly modified the perceived-distance position of the card from one reduced-cue situation to another, when the same subjects were used in the several situations.

TABLE 6
SIGNIFICANCE OF THE DIFFERENCES BETWEEN PERCEIVED-DISTANCE SCORES FROM ALL PRESENTATIONS OF THE REDUCED-CUE SITUATIONS

Test	Qualitative Differences	Difference in Feet	t	$df = \frac{p^a}{1/237}$
1. $P_B - P_A > 0$	Angle	4.8	6.24	$< .01$
2. $P_D - P_C > 0$	Angle	5.6	7.29	$< .01$
3. $P_C - P_B > 0$	Distance	1.5	1.96	$< .025$
4. $P_C - P_A > 0$	Angle and Distance	6.3	8.20	$< .01$
5. $P_D - P_B > 0$	Angle and Distance	7.1	9.27	$< .01$
6. $P_D - P_A > 0$	Angle and Distance	11.9	15.50	$< .01$
7. $(P_D - P_A) - (P_C - P_B) > 0$	Angle Corrected for Distance	10.4	9.57	$< .01$

^a Single-tailed probability of occurrence.

The analysis of the results from the first-presented reduced-cue situations (Tables 2 and 3) gave no indication that absolute-distance perception had occurred as a function of the retinal size of the playing card. The analysis did suggest, however, that some perception of absolute distance had occurred as a function of distance cues other than those of retinal size. These cues would operate to make the physical and apparent distances of the playing cards coincide. The degree to which the physical and apparent distances are not in agreement indicates the extent to which these residual cues of absolute physical distance were unimportant in producing perceptions of absolute distance in the reduced-cue situations. This can be evaluated by comparing the perceived-distance scores of Table 1 with the actual physical distances of the cards. It will be observed from Table 1 that, when the cards were at 10 feet, the perceived-distance scores were greater than 20 feet. That the perceived-distance scores are not proportional to the physical distances in the first presentations of the reduced-cue situations is also suggested by Table 1. Computing the ratio of the largest to the smallest perceived-distance score in Table 1 gives a value of 1.4. The corresponding ratio of the physical distances is 2.0. It may be concluded that the residual cues to the physical distance of the playing cards were not very effective in producing perceived absolute distance in the reduced-cue situations. The perception of absolute distance in the reduced-cue situations will be regarded as having been largely indefinite. Consequently, it will be considered that the subject threw to what was essentially an indifferent distance when first presented with a reduced-cue situation. This throwing distance was not indifferent,

however, in that a throwing force was used which was definite, and which probably was peculiar to the particular subject. Later, the same subject, upon being presented with a second reduced-cue situation, modified the throwing force depending upon whether the card in the second presentation appeared to be more or less distant than that in the first presentation. Consider the case in which a subject was presented with the reduced-cue situations in the order $A_1 B_2 C_3 D_4$, where the subscripts refer to the order of presentation. Upon being presented with A_1 , the darts would be thrown to some indifferent distance. When the same subject was presented with B_2 , he behaved as though he remembered both the retinal subtense of A_1 , in relation to that of B_2 , and the force of the throwing he had used in A_1 . Since the card in B_2 was retinally smaller than that in A_1 , the card in B_2 was judged to be more distant than the previously indefinitely perceived distance of the card in A_1 , and the force of the throwing was increased. The throwing on C_3 and D_4 for the same subject would be determined in a similar manner. For the order of presentation $A_1 B_2 C_3 D_4$, the relative magnitude of the throwing would be $T_{A_1} < T_{B_2} = T_{C_3} < T_{D_4}$. *Since the perception of an absolute distance as a function of retinal subtense was not demonstrated by the analysis of the results from the first-presented reduced-cue situations, the perception of distance which occurred between successively presented reduced-cue situations as a function of retinal subtense is considered to be relative-distance perception.* The results indicate that relative-distance perception, as some function of relative retinal subtense, occurred between the successively presented reduced-cue situations.

If relative-distance perception occurred as a function of the relative retinal subtense of the playing cards, it would be expected that the order of magnitude of the throwing scores (or perceptual scores) in the successively presented reduced-cue situations would be independent of the order of presentation. This can be demonstrated by considering the case in which a subject was successively presented with the reduced-cue situations in the order $D_1 C_2 B_3 A_4$. After making a perceptually indifferent throwing score on D_1 , the subject, when presented with C_2 , should produce a throwing score which was less than that which resulted from D_1 . This would be a consequence of the retinal subtense of the card in C_2 being greater than that of the card in D_1 . Similarly, A_4 should have the smallest throwing score of the four situations. The relative magnitude of the throwing scores would be $T_{A_4} < T_{B_3} = T_{C_2} < T_{D_1}$. As discussed above, this is the same order of relative magnitude of throwing scores which would be expected from the order of presentation $A_1 B_2 C_3 D_4$. Continuing this, it will be seen that the order of magnitude of the throwing (or perceptual scores) would be independent of the order of presentation of the reduced-cue situations. It is this expectation which, in the comparative absence of absolute-distance perception in the first-presented situations, permits the analysis of the results from all the presentations of the reduced-cue situations to be interpreted.

3. *The Independence of Absolute- and Relative-Distance Perception*

In the introduction, it was suggested that the characteristic which discriminates between apparent absolute and apparent relative distance is the refer-

ence point with respect to which the perception occurs. This reference point is some function of the position of the observer for absolute-distance perception, but is independent of the position of the observer for relative-distance perception. It has been pointed out that experimental results which have been attributed to an absolute-distance perception might be explained by a relative-distance perception. Conversely, the difference between two absolute-distance perceptions might be considered, mistakenly, to be a perception of relative distance, even though the perceptual reference point in this case was not independent of the position of the observer. The results from the reduced-cue situations do not support the hypothesis that perceived absolute distance occurred in the reduced-cue situations as a function of the retinal size of the playing card. But the results do indicate that relative-distance perception occurred as a consequence of changes in retinal size. This suggests that, as a function of retinal size, relative-distance perceptions might have occurred in the reduced-cue situations without a concomitant occurrence of absolute-distance perceptions. This offers the possibility that an apparent relative distance between two objects can occur independently of the perception of the distance of either of these objects. Thus, the definition of relative-distance perception used in this study may parallel a real perceptual ability. A test of this possibility would be a test in which an affirmative answer simultaneously would depend upon the absence of perceived absolute distance and the presence of perceived relative distance.

While the order of magnitude of the perceptual (or throwing) scores should

be independent of the sequence of presentation of the reduced-cue situations, this is not the case with the absolute magnitude of the perceptual (or throwing) scores. Consider, for example, the sequence of presentation $A_1 B_2$ as compared with the sequence of presentation $D_1 B_2$ and assume, for the moment, that all distance cues except that of retinal size had been completely eliminated. Under the condition that no absolute-distance perception occurred in the reduced-cue situations, P_{A_1} and P_{D_1} should not differ significantly from each other. With the sequence $A_1 B_2$, P_{B_2} as a consequence of the smaller retinal size in B_2 than in A_1 should be greater than P_{A_1} . With the sequence $D_1 B_2$, P_{B_2} should be less than P_{D_1} . Therefore, P_{B_2} with the sequence of presentation $A_1 B_2$ should be greater than P_{B_2} with the sequence $D_1 B_2$. This is based upon the assumptions that (a) absolute-distance perception did not occur with A_1 and D_1 , but (b) relative-distance perception occurred between successively presented situations. Similarly, P_{C_2} with the sequence of presentation $A_1 C_2$ should be greater than P_{C_2} with the sequence $D_1 C_2$. Both B_2 and C_2 should have resulted in larger scores when preceded by A_1 than by D_1 .

Actually, as shown in Table 3, test 6, P_{D_1} was significantly larger than P_{A_1} . This is attributed to the effect of residual cues of absolute distance, other than the cue of retinal size. To the extent that these cues to the physical distance of the target in either A and D or B and C were present, B_2 and C_2 should not have resulted in larger scores when preceded by A_1 than by D_1 . Larger scores in B_2 and C_2 , as a consequence of being preceded by A_1 rather than by D_1 , indicate that relative retinal size can produce a per-

TABLE 7
AVERAGE PERCEPTUAL SCORES IN FEET FROM
SITUATIONS B_2 AND C_2 , DEPENDING UPON
WHETHER THE PRECEDING SITUATION
WAS A_1 OR D_1

Prior Situations	Results from B_2		Results from C_2	
	Mean	SD	Mean	SD
A_1	25.6	7.6	26.6	6.1
D_1	17.2	9.7	19.2	6.1
Difference	8.4		7.4	
Average Difference		7.9		

ception of the relative distance between two objects, which is independent of a perception of the absolute distance of either.

In this study there were seven subjects who had the sequence $A_1 B_2$; seven who had the sequence $D_1 B_2$; seven who had the sequence $A_1 C_2$; and seven who had the sequence $D_1 C_2$. Table 7 gives the average perceptual scores from B_2 and C_2 , depending upon whether the preceding situation was A_1 or D_1 . The average difference of 7.9 feet is the difference between the mean of 14 scores on B_2 or C_2 , when the preceding situation was A_1 , and the mean of 14 scores on B_2 or C_2 , when the preceding situation was D_1 . It is the average amount by which the perceptual scores from B_2 and C_2 were larger as a consequence of being preceded by A_1 rather than by D_1 . This average difference was significant beyond the .01 level of confidence ($t = 2.66$, $df = 26$) with the single-tailed t test. This indicates that the presence of relative-distance perception as a function of the size of a familiar object can occur independently of the occurrence of absolute-distance perception.

IV. DISCUSSION

A. The Full-Cue Situations

It is reasonable to expect the perception of absolute distance to occur in the full-cue situa-

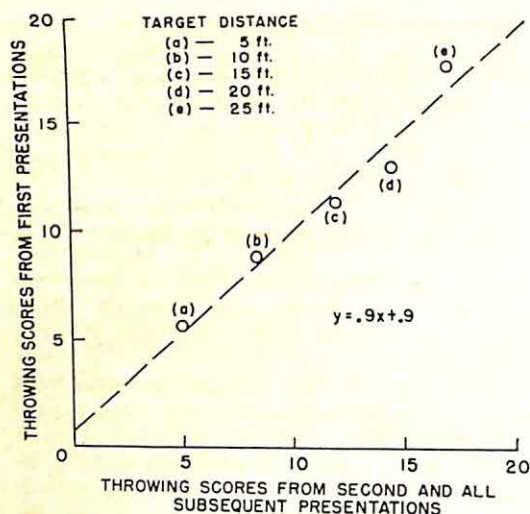


FIG. 3. Relation between the mean throwing scores in feet from the first and all subsequent presentations of the full-cue situations.

tions. It can be demonstrated that, unlike the results from the first presentations of the reduced-cue situations, the throwing in the first presentations of the full-cue situations was not largely indifferent. This can be demonstrated by comparing the average throwing scores from the first presentations of the full-cue situations with the average throwing scores from the remaining presentations of the full-cue situations. This is shown in Fig. 3 where this relationship is plotted for the five distances of the playing card. It will be seen that the data reasonably fit a straight line. A straight line was fitted to these data by the method of least squares. The equation of this line is given in Fig. 3. The similarity of this equation to the equation $y = x$ indicates that the throwing scores in the first and remaining presentations of the full-cue situations were essentially the same.

B. Experimental Precision

This study produced no evidence that any relation occurred between the retinal size of the familiar object and the perception of absolute distance. But it cannot be concluded that no such relation existed, since this conclusion would require the acceptance of the null hypothesis. This raises the problem of the experimental precision with which the relation between perceived absolute distance and the retinal size of a familiar object has been investigated. This experimental precision can be considered either from the standard errors involved in the study, or from the kind of perceptual discriminations which the study was able to demonstrate.

One possible estimate of experimental precision is the error term which was used to test the statistical significance of the average score difference between first presentations of the reduced-cue situations. This error term of 2.9 feet would suggest that a perceived difference between situations of 4.8 feet (1.645×2.9 feet) would have to occur before a statistically significant mean difference would be present with the one-tailed t test. On rational grounds, however, this is not a good estimate of the experimental precision with which familiar size was tested as a cue to absolute distance. The error term from the first-presented reduced-cue situations would vary depending upon the precision with which the perception of absolute distance occurred. For example, with no perception of absolute distance, the variability between throwing scores from the first presentations of the reduced-cue situations would be large, since the throwing would not be controlled by a visually perceived distance. The problem is to determine an error term which reflects the precision of measurement which would occur if the perception of absolute distance were definite in the reduced-cue situations. In the full-cue situations it was assumed that the perception of absolute distance was accurate. The standard error, calculated from the average between-subject variability in the full-cue situations, is a possible measure of the precision with which the existence of a relation between perceived absolute distance and retinal size was tested in the reduced-cue situations. This was calculated by computing the standard deviation of the subjects' scores in each of the full-cue situations, for each of the four groups of 20 subjects who had previously been used in the first presentations of the reduced-cue situations. The resulting average standard deviation was 2.0 feet. Using this average standard deviation, the standard error of the mean difference between the results from different groups of 20 subjects is .7 feet. With this standard error, a mean difference of 1.1 feet would be significant at the .05 level of confidence, with the one-tailed t test. This indicates that with two different groups of 20 subjects, with each group presented with a different reduced-cue situation, a mean difference between groups of 1.1 feet in perceived absolute distance would be significantly differentiated by the throwing procedure at the .05 level of confidence. As noted previously, the error term calculated from the reduced-cue situations is probably too large to be considered an estimate of the precision with which the relation between the retinal size of a familiar object and perceived absolute distance was tested. The estimate from the full-cue situations, however, is possibly too small for this purpose, since it contains the assumption that, if the perception of absolute distance occurred

in the reduced-cue situations, the variability between subjects attributable to this perception would be no greater than the variability which occurred in the full-cue situations. To the degree that additional perceptual variability can be tolerated without rejecting the possibility of some general relation between retinal size and perceived absolute distance, this estimate of experimental precision is too small. Therefore, the experimental precision with which the occurrence of the relation between the retinal size of a familiar object and perceived absolute distance was tested lies somewhere between these two possible estimates, i.e., somewhere within the range of approximately 1 to 5 feet.

It is possible to consider conclusions which reject the null hypothesis with respect to any particular relation between retinal size and the perception of absolute distance. One such relation is that given by the size-distance hypothesis. The perceived distances predicted from the size-distance hypothesis for the reduced-cue situations are given in Table 1. The perceptual scores of Table 1 are logically the values to compare with the corresponding predicted distances. The error term from the analysis of variance of these perceived distance scores (2.9 feet) can be used to calculate the standard error, and thus the statistical significance of the differences between the perceptual scores and the corresponding predicted scores. If this is done, it is found that all of these differences are significant beyond the .01 level of confidence, by the double-tailed *t* test. This supports the conclusion that the magnitudes of the obtained perceptual scores are not in agreement with the values predicted from the size-distance hypothesis.

Any statistical significance which occurred between the average results of the reduced-cue situations is evidence that the throwing method can measure perceived depth in these situations. Statistical significance was present between the average results from the successively presented reduced-cue situations, indicating the occurrence of relative-distance perception. The standard error used to test the significance of the mean difference between any two such situations was .8 feet for the perceived-distance scores. This is much smaller than the corresponding error term of 2.9 feet from the perceived-distance scores with the first-presented reduced-cue situations. The experimental precision of the test for relative-distance perception was greater than that for absolute-distance perception. But it was also found that the perceived-distance score from situation B or C varied, depending upon whether B or C was preceded by A or D (see Table 7). The standard error for the test of significance of this comparison was 3.0 feet. This is of the same order of magnitude as the standard error of pairs of mean differences from the first-pre-

sented reduced-cue situations (2.9 feet). This indicates that, with standard errors of this magnitude, a statistically significant mean difference can appear, when the basis for expecting this difference is that relative—but not absolute—distance perception is some function of the retinal size of familiar objects.

C. Perception of Distance

The first time a reduced-cue situation was presented to a subject, the situation was a unique presentation in that the subject probably had never previously encountered a situation identical to it. But the subject had encountered normal-sized playing cards prior to the experiment. (This prior experience is obviously necessary in order for the playing card to be a familiar object.) Since these prior experiential situations were usually full-cue situations, the absolute distance of the object was undoubtedly perceived. It might be expected that, upon being presented with a particular retinal subtense of a playing card in a reduced-cue situation, the subject would make a distance judgment relative to this prior experience. If such a result had been demonstrated, it would have been difficult to decide whether the distance judgment should be termed absolute or relative. Perhaps the reason that such a result did not appear in this study is that too long an average time interval occurred between the prior experience and the experiment. It will be recalled that, with the relative localization demonstrated in the reduced-cue situations, only a four-minute interval separated the successive presentations.

This emphasizes the difficulty in demonstrating absolute-distance perception without the possibility that relative-distance perception had produced the results. But, as has been indicated by this study, a relative-distance perception as a function of relative retinal size can occur independently of the occurrence of absolute-distance perception.

This study does not suggest that retinal size has no part in the perception of absolute distance. In many situations, the perception of absolute distance may at least partially consist of a series of successive relative-distance perceptions. In this sense, a series of relative retinal sizes may be a significant factor in determining the perception of absolute distance. Also, there is no evidence that, given the proper training, the subjects could not have learned to use the absolute retinal size of a familiar object as a cue to determine its absolute distance. But, a distinction can be made between an immediate distance perception and an inferential estimate of distance (2, p. 342). Therefore, whether such learning would result in the ability to make an inferential estimate of absolute distance or would

produce an absolute-distance perception would remain to be investigated.

V. SUMMARY AND CONCLUSIONS

An experiment was conducted to determine whether the retinal subtense of a familiar object could act as a determiner of its apparent absolute distance. To test this question, it was required that each subject be presented with only one object in an otherwise homogeneous field, and that different subjects be used with the different retinal sizes of the same familiar object. Apparent distance was measured by having the subjects throw darts to the apparent distance-position of a playing card. Neither the trajectory, nor the place of impact of the darts was visible to the subjects. Two kinds of situations were used. In one, called the reduced-cue situations, the purpose was to eliminate all distance cues but retinal size. In the

other, called the full-cue situations, a variety of distance cues were present. The results from the full-cue situations were used to change throwing scores from the reduced-cue situations to perceived-distance scores. The analysis of the average differences in the scores between the *first* presentations of different reduced-cue situations revealed no evidence for the presence of absolute-distance perception as a function of retinal size. However, between the *successively* presented reduced-cue situations, a change in the magnitude of the scores as a consequence of a change in retinal size did occur. This was interpreted to be the result of relative- not absolute-distance perception. It seems that relative-distance perception as a function of the relative size of familiar objects can occur without the concomitant occurrence of absolute-distance perception.

REFERENCES

1. HASTORF, A. H. The influence of suggestion on the relationship between stimulus size and perceived distance. *J. Psychol.*, 1950, 29, 195-217.
2. HOCHBERG, C. B., & HOCHBERG, J. E. Familiar size and subception in perceived depth. *J. Psychol.*, 1953, 36, 341-345.
3. ITTELSON, W. H. Size as a cue to distance: static localization. *Amer. J. Psychol.*, 1951, 64, 54-67.
4. ITTELSON, W. H., & AMES, A., JR. Accommodation, convergence and their relation to apparent distance. *J. Psychol.*, 1950, 30, 43-62.
5. KILPATRICK, F. P., & ITTELSON, W. H. The size-distance invariance hypothesis. *Psychol. Rev.*, 1953, 60, 223-231.
6. SCHLOSBERG, H. A note on depth perception, size constancy, and related topics. *Psychol. Rev.*, 1950, 57, 314-317.

(Accepted for publication March 2, 1957)

Psychological Monographs: General and Applied

The Effect of Three Teaching Methods on Achievement and Motivational Outcomes in a How-to-Study Course

JOHN D. KRUMBOLTZ AND WILLIAM W. FARQUHAR

*Michigan State University¹*I. THE PROBLEM¹

INSTRUCTION in study skills has been a part of the American college scene since the early 1920's. Since that time much effort has been expended developing techniques of skill and evaluating their effectiveness, but little attention has been given to the role of the teacher in teaching the skills. The present investigation is an attempt to analyze several methods of instruction as they relate to student outcomes in a how-to-study course.

Two broad classifications of outcomes are considered. The first is termed "achievement outcomes," and it is defined in terms of the specific learnings which take place during the course of instruction. The learnings involve specific knowledge about effective methods of study, knowledge of research studies and their outcomes that cast light on study skills, knowledge of human characteristics that may help or hinder studying, and specific facts about reading, writing, outlining, note-making, testing, and concentrating that should help in study problems. Knowledge of this nature is important in gaining an understanding of the purpose and techniques of effective study, but in and of itself does not guarantee practical application of the techniques learned. Therefore, it is also important to measure the extent to which students adopt the methods they learn into their own study habits, and the extent to which they form attitudes conducive to good study. Thus, knowledge and practice are two indispensable measures of student achievement in a how-to-study course.

The second classification of outcomes may be termed the "motivational outcomes"—an area

which has been virtually ignored in experimental studies on the outcomes of teaching. Educators have not been unaware of the problem of motivation; the educational literature contains many references to the importance of motivating students. Many techniques and devices have been suggested in an attempt to raise the motivational level of students. However, almost no experimentation has been done to determine the actual value of the suggestions. For the purposes of this study, one particular motive is under consideration—the need to achieve. Need to achieve may be considered as an internal drive toward a specific action directed toward achieving what is called "success" in the American culture. Achieving success involves being able to gain the recognition of one's fellow man; it involves striving to compete against some standard of excellence. In short, it may be called "drive" or "ambition." The achievement motive is not confined to one field of endeavor. The actual task is unimportant; it is the individual's attitude or internal state of affairs impelling him to be successful which is important.

Purpose

The primary purpose of the present investigation is to study both the achievement outcomes and the motivational outcomes associated with different methods of teaching in a how-to-study course. We are concerned with identifying instructional methods that are most effective in improving the students' knowledge of effective study methods, in encouraging the actual application of these methods, and in increasing the students' level of achievement motivation.

Two important subsidiary problems involve the question of whether the method that is best for one type of student is also best for another. First, what influence does preference for a certain

¹This monograph is a compilation of two Ph.D. theses completed at the University of Minnesota. The authors are indebted to their major advisors, C. Gilbert Wrenn and Willis E. Dugan, for their advice and assistance in this research.



type of instruction have on achievement and motivational outcomes? For example, do students achieve better when they receive the type of instruction they prefer, or do they perhaps learn more under an instructional method they do not like? Second, what relationship exists between student ability-level and type of instruction received? For example, do bright students achieve better under one type of instruction while the less bright students improve more under another teaching method? The various interactions between teaching method and student preference for method, and between teaching method and student ability-level are tested in this investigation. The specific null hypotheses are outlined in Part II.

Unique Aspects of the Study

Birney and McKeachie (2), in a 1955 review of the literature on teaching methods in psychology, defined two principal types of teaching methods—an instructor-centered and a student-centered method. Most of the studies cited by Birney and McKeachie consist of a comparison of the effectiveness of these two methods. Why is it necessary to use only these extreme methods? Is it not conceivable that a combination of methods might be more effective than strict adherence to any one? The present experiment makes use of an eclectic teaching method, an attempt to combine the advantages of both of the extreme methods. At the same time an instructor-centered and a student-centered method are also compared. The methods are defined in Part III.

Another important feature of this investigation is that it measures motivational outcomes in addition to achievement outcomes. To the best of the authors' knowledge, no studies have yet been performed measuring directly motivational outcomes associated with different teaching methods. The reason for this lack has probably been that adequate measures of motivation have not been available. It is only within the last few years that relatively valid measures of motivation have been developed. The measures used in the present investigation are described in Part IV.

The design of the experiment constitutes an important improvement over the design of many other studies in this area. Farquhar (6) has criti-

cally reviewed the literature on teaching methods and has pointed out many of the methodological faults underlying previous experiments. The present investigation seeks to overcome many of the design limitations of previous studies in order to secure more defensible interpretations of results. The details of the design are outlined in Part II.

II. THE DESIGN OF THE EXPERIMENT

The study is designed to investigate differences in achievement and motivational outcomes which may be associated with three teaching methods. As a self-contained experiment, it embodies three principles essential to modern research: randomization, replication, and control (12).

The Basic Plan

The teaching methods used in this study are instructor-centered, student-centered, and eclectic. Two instructors at two different times teach six sections of the how-to-study course using all three methods. Instructor A teaches by all three methods during the morning; instructor B teaches by all three methods during the afternoon hours. The relative contribution of the different times and the different instructors to experimental outcomes cannot be separated within the limits of this design. The combined effect will henceforth be referred to as time-instructor.

The design of the experiment is summarized in Table 1.

TABLE 1
SUMMARY OF THE BASIC EXPERIMENTAL DESIGN

Time-Instructor	Method		
	Instructor-Centered	Eclectic	Student-Centered
Instructor A: Morning	Section 1	Section 2	Section 3
Instructor B: Afternoon	Section 4	Section 5	Section 6

Complete randomization was not possible because of administrative difficulties. Students were randomly placed among the three teaching methods, but not between time-instructors. It is necessary to isolate the differences between the effect of time-instructors in the analysis, but this effect is not central to the purpose of the study. The primary problem is to detect differences in outcomes among methods. This randomization procedure provides the basis for obtaining a valid estimate of experimental error so that a probability statement about method differences based on a known distribution can be made.

Replication is accomplished by having each method taught by two instructors rather than one. This not only permits a valid estimate of experimental error, but also tends to diminish experimental error.

Each method provides a control on each of the others. No attempt is made to provide a section which is given no instruction in how to study. No one section is designated as the control section or as the experimental section. Each section is both a control and an experimental section in relation to each of the other sections.

Primary Null Hypotheses

Four criterion measures are to be used in the main analysis and a complete description of these instruments is found in Part IV. With these measures the following null hypotheses are to be tested:

1. No differences exist among mean scores of the criterion instruments attributable to different methods of teaching.
2. No differences exist between mean scores of the criterion instruments attributable to different time-instructors.
3. No interaction effect exists between teaching method and time-instructor.

Subsidiary Null Hypotheses

From the many possible subsidiary problems growing out of a study of this nature, two seem to be especially pertinent. One involves an attempt to discover relationships between preference for certain instructional methods and different outcomes associated with the actual teaching by these methods. The other is concerned with the relationship

of students' ability level to different outcomes of these methods.

The following procedure is utilized in making the first subsidiary analysis: The students taught by identical methods (though by different instructors) are pooled and then split according to their preference for certain teaching methods. Preference is measured by the Preferred Instructor Characteristics Scale (PICS) which is fully described in Part IV. The design is represented in Table 2.

The following subsidiary null hypotheses are to be tested:

1. No differences exist among mean scores of the criterion instruments attributable to different instructional methods.
2. No differences exist among mean scores of the criterion instruments attributable to preference for teaching methods.
3. No interaction effect exists between teaching method and preference for instructional method.

The main concern of this design is with the third hypothesis—the test for interaction. From this hypothesis it will be possible to determine whether students with a certain preference are more motivated by one instructional technique than another.

A similar procedure is followed in analyzing ability-level relationships. The students taught by identical methods are

TABLE 2

SUMMARY OF DESIGN FOR SUBSIDIARY ANALYSIS OF PREFERENCE FOR INSTRUCTIONAL METHOD

Preference for:	Actual Method of Instruction		
	Instructor-Centered	Eclectic	Student-Centered
Instructor-Centered	(Scores on each of the criterion tests are entered in the nine cells of this table.)		
Eclectic			
Student-Centered			

TABLE 3

SUMMARY OF DESIGN FOR SUBSIDIARY ANALYSIS
ON ABILITY LEVEL

Ability Level	Actual Method of Instruction		
	Instructor-Centered	Eclectic	Student-Centered
High	(Scores on each of the criterion tests are entered in the nine cells of this table.)		
Third			
Medium			
Third			
Low			
Third			

again pooled and this time split into thirds on the basis of their scholastic aptitude level as measured by the American Council on Education Psychological Examination (ACE). The design is summarized in Table 3.

The following null hypotheses are to be tested:

1. No differences exist among mean scores of the criterion instruments attributable to different teaching methods.

2. No differences exist among mean scores of the criterion instruments attributable to different ability levels.

3. No interaction effect exists between teaching method and ability level.

Here, as in the first subsidiary analysis, the primary concern is with the interaction hypothesis—the indication of whether students of certain ability-levels are more affected by one instructional technique than another.

The procedure of pooling students taught by identical methods but different time-instructors is justifiable in the subsidiary analysis if there are no significant time-instructor differences and no significant interaction between time-instructor and method in the primary analysis. If the preceding conditions are not met in the primary analysis, the data can be analyzed by a three-way analysis of variance and covariance, isolating the effects of time-instructor, method, and ability level or preference for method variables.

Statistical Analysis

Analysis of variance and analysis of

covariance are the fundamental statistical tools used in the present investigation. It is possible by the use of these tools to isolate the sum of the squares associated with each experimental variable and test its significance statistically.

Certain assumptions are involved in the use of these tools.

1. The observations should be normally distributed within each cell. This assumption of normality is seldom tested. As Johnson states, "This assumption, however, is not especially important" (12). It is not tested in the present investigation.

2. Variance within cells should be equal. This is an important assumption and is tested by the L_1 Test (see reference 12).

3. Observations should be randomly selected. Details of the present randomization procedure are found in Part III.

4. In covariance problems the linear regression coefficients of posttest scores on pretest scores should be equal within cells. The test of this assumption is provided by a test of the homogeneity of regression coefficients (see reference 10).

Each of the four criterion tests is administered before and after the instructional period. The results of the *first administration* are hereafter referred to as *pretest* scores and the results of the *second administration* as *posttest* scores. By this technique, it is possible to adjust statistically differences in posttest scores for initial differences in pretest scores through analysis of covariance.

The number of students falling in each cell is not equal. Therefore, a special computational technique is required in a two-way analysis of variance, if an exact test of significance is desired. Such an exact test, involving a least-square analysis solved by pivotal condensation (4), is employed in the present investigation.

The level of significance for rejecting the primary and subsidiary null hypotheses is arbitrarily set at the five per cent level. For tests of assumptions the one per cent level is set.

III. EXPERIMENTAL PROCEDURES

This part is concerned with outlining in detail the experimental procedures specified in the design. The selection of

a random sample, the specific definitions of the instructional situation and methods, and the various checks on the qualitative differences among methods and the consistency of instruction between instructors are developed in this order.

Selection of the Sample

One assumption underlying the use of the statistical tools of this investigation is randomness of observations. Before the experiment actually started, it was therefore necessary to specify the population accurately, and to devise a procedure to assign this population randomly to the three teaching methods under investigation.

The Population

The population for this study consists of all University of Minnesota students who elected to take Personal Orientation 1, How to Study, in the winter quarter of 1955. This population consists of both male and female students from five colleges of the University: the College of Science, Literature, and the Arts, the Institute of Agriculture, the Institute of Technology, the College of Education, and the General College. Most of these students are freshmen and sophomores, but a few are upperclassmen.

In no way can this population be considered a random sample of the entire college and university population of the United States, nor can it be considered even representative of the University of Minnesota. Students enroll in a how-to-study course for a variety of reasons. Most feel that they need help with their study problems and are motivated to the point where they want to take action. No conclusions about this population can justifiably be generalized to other dissimilar populations.

Randomization Procedure

Students desiring to register for Personal Orientation 1, How to Study, given in the winter quarter of 1955, were required to have the written consent of the instructor. Obtaining permission of the instructor is not the usual procedure with this course, but it was considered necessary as an aid in the randomization procedure.

The purpose of the randomization procedure was to place students at random in one of the three teaching methods. No attempt was made to place students randomly between time-instructors. The procedure was designed so that students could have their choice of a morning or afternoon class and the corresponding instructor. But once that decision was reached, the one section out of the three possible sections of that time-instructor was determined randomly.

When a student appeared for registration, he was given a choice of two sections which had been predetermined in a random manner except that one was always from the morning sections and the other from the afternoon sections. If he were able to accept one of these two sections, he was entered as a random student for that section. The next student to appear for registration was given his choice of the next pair of randomly determined section numbers.

If, for any one of a number of acceptable reasons, a student was unable to work the randomly assigned section into his schedule, he was allowed to sign up for the section he desired, but was designated a nonrandom student in that section. The next student to appear was then given the choice of the same two section numbers which had been unacceptable to the previous student.

The procedure continued until one of the sections was filled to capacity (30 students). Thereafter, no more random assignments were made in the remaining two sections for that time-instructor. The succeeding registrants in those two sections were all designated as nonrandom. Random assignment continued in the other time-instructor sections until one of these sections was filled. All of the remaining registrants were then designated as nonrandom. At the end of the procedure 192 students had registered. Of these, 121, or 63 per cent, were randomly assigned. The remaining 71 students, or 37 per cent, were nonrandomly assigned.

Characteristics of the Population

During the winter quarter in which instruc-

TABLE 4
NUMBER OF STUDENTS FALLING IN EACH COLLEGE CLASS AND SECTION

Class	Random or Nonrandom	Section						Total
		Time-Instructor A			Time-Instructor B			
		1	2	3	4	5	6	
Freshmen	Random Nonrandom	13 6	16 6	17 9	13 8	17 4	8 8	84 41
Sophomores	Random Nonrandom	6 1	3 2	3 2	2 2	3 3	6 7	23 17
Juniors	Random Nonrandom	1 0	0 0	1 0	2 2	0 2	1 0	5 4
Seniors	Random Nonrandom	0 2	0 0	0 0	0 1	0 0	0 0	0 3
Total	Random Nonrandom	20 9	19 8	21 11	17 13	20 9	15 15	112 65

tion took place, fifteen students dropped from the course. No particular disproportion appears among the sections, between time-instructors, or between the random and nonrandom groups. It seems safe to conclude that the distribution of dropouts introduces little bias into the results.

The remaining students consist of 112 random and 65 nonrandom students. The distribution of these students by college class and by section appears in Table 4. About 75 per cent of the randomly placed students are freshmen, 21 per cent are sophomores, 4 per cent are juniors, and zero per cent are seniors. In general, the placement of the nonrandom students was similar to that of the random.

The proportion of random students to the total enrollment in each of the six sections tends to be approximately the same. The range is from 50 per cent in section 6 to 70 per cent in section 2. The proportion of random students

in the entire group is 63 per cent.

Distribution of students by sex is presented in Table 5. It is evident that the proportion of each sex within sections and between random and nonrandom groups remains reasonably constant. However, as an additional check on this factor, separate analyses are reported in Part V for the males only and for the males and females combined.

The distribution of students by college is summarized in Table 6. About 75 per cent of the randomly assigned students are registered in the College of Science, Literature, and the Arts, 12 per cent in the College of Education, 3 per cent in the Institute of Technology, 4 per cent in the Institute of Agriculture, and 6 per cent in the General College. The proportion of nonrandom students from each of these colleges is similar, considering the small frequencies involved.

TABLE 5
NUMBER OF STUDENTS OF EACH SEX IN EACH SECTION

Sex	Random or Nonrandom	Section						Total
		Time-Instructor A			Time-Instructor B			
		1	2	3	4	5	6	
Male	Random	17	16	19	10	12	9	83
	Nonrandom	7	7	9	11	7	12	53
Female	Random	3	3	2	7	8	6	29
	Nonrandom	2	1	2	2	2	3	12
Total	Male	24	23	28	21	19	21	136
	Female	5	4	4	9	10	9	41

TABLE 6
NUMBER OF STUDENTS FROM EACH COLLEGE OF THE UNIVERSITY, BY SECTION

NUMBER OF STUDENTS FROM EACH								
College	R = Random NR = Nonrandom	Section						Total
		Time-Instructor A			Time-Instructor B			
		1	2	3	4	5	6	
Science, Literature, and Arts	R	14	14	19	11	14	12	84
	NR	6	7	10	11	8	13	55
Education	R	1	3	0	5	3	1	13
	NR	2	0	0	1	1	1	5
Institute of Tech- nology	R	1	0	1	0	0	1	3
	NR	0	0	0	0	1	0	1
Institute of Agricul- ture	R	2	0	1	1	1	0	5
	NR	0	0	0	0	0	1	1
General College	R	2	2	0	0	2	1	7
	NR	1	1	1	0	0	0	3
Total	R	20	19	21	17	20	15	112
	NR	9	8	11	13	9	15	65

An accurate attendance record was kept throughout the quarter. It was important to check this factor in order to ascertain whether the students from different sections had an equal exposure to the instructional method. Out of a total of 29 class periods during the quarter, the random students missed an average of 2.07 class periods, or 7.1 per cent, while the nonrandom students missed an average of 2.72 class periods, or 9.4 per cent. The average number of class periods missed for the random students of each section remained relatively constant over all six sections. The range was from 1.65 class periods (5.7 per cent) in sections 4 and 5 to 2.79 class periods missed (9.6 per cent) in section 2. The

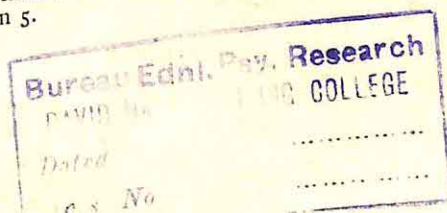
absences were few enough in any section and relatively constant enough among sections to justify the conclusion that this factor introduces little bias into the results.

Do the sections vary in terms of their average scholastic aptitude? Records of the American Council on Education Psychological Examination (ACE) were obtained from the Student Counseling Bureau. Scores on different forms of this examination were converted to equivalent raw scores on the 1952 form by means of conversion tables provided by the Student Counseling Bureau. A summary of the mean raw scores of random students for each section appears in Table 7.

TABLE 7
MEAN RAW SCORE ON THE AMERICAN COUNCIL ON EDUCATION
PSYCHOLOGICAL EXAMINATION FOR EACH SECTION
RANDOMLY PLACED STUDENTS ONLY

Sex	Time Instructor	Method			Total
		Instructor Centered	Eclectic	Student- Centered	
Males and Females Combined	A B Total	98.75	107.95	103.86	104.05
		96.69	98.05 ^a	94.87	96.69 ^a
		97.83	102.87 ^a	100.11	100.27 ^a
Males Only	A B Total	99.76	111.06	103.32	104.54
		94.89	99.50 ^a	89.89	95.23 ^a
		98.08	106.11 ^a	99.00	101.13 ^a

^a The number of students in these cells is one less than would be expected because no ACE score was available for one randomly placed male in Section 5.



An examination of Table 7 reveals a tendency for the sections of time-instructor A to score higher than the sections of time-instructor B, while scores in sections taught by the same method seem to be relatively more homogeneous. An exact test of significance using analysis of variance was made to investigate the statistical significance of these differences.

For the males and females combined, differences in mean scores between time-instructors and among methods do not reach significance. However, when males only are analyzed, the difference between time-instructors becomes significant at the five per cent level, but differences among methods remain insignificant. The interactions are not significant in either case.

This significant difference between time-instructors could possibly be a result of the manner of randomization. Students were given a choice of a certain time-instructor but were randomly assigned to methods. Possibly some selective factor operated in making this choice to create the significant difference. The important fact, however, is that no significant differences exist among methods. Apparently the randomization procedure was successful in distributing students among the three methods so that approximately equal means and variances resulted.

To compare the average ACE standing of the random and nonrandom students, Student's *t* test, as reported by Johnson (12, p. 72), is used. The 61 nonrandom students on whom ACE scores are available have a mean raw score of 103.23; the 111 random students have a mean of 100.34. The variances are approximately equal ($F = 1.35$). The resulting *t* value is .917 and is nonsignificant. It may be concluded that the two groups are approximately equal in scholastic ability.

The Nonrandom Students

The 65 students who could not be randomly assigned are not included in the analysis of results. The data presented in the preceding section show that they are not essentially different from the randomly placed students on any of the characteristics examined. However, it is impossible to examine the differences between the two groups on every variable that might have any relevance, and no attempt in that direction has been made.

In order to meet the assumptions of the statistical tools, it was judged best to confine the analysis only to the randomly placed students. The possible influence of the nonrandom students on the random students in each section has been considered, but the subjective impression of the instructors is that they exerted no influence out of proportion to their number. The students themselves did not know whether they

had been randomly or nonrandomly assigned, and this concept was never discussed with them.

The Instructional Situation

For a more complete understanding of the experiment, it is important to consider the setting, the particular course involved, and the instructors themselves. This section is devoted to a brief description of these factors.

The Setting

All classes met in Room 307, Johnston Hall, on the University of Minnesota Minneapolis campus. The room was equipped with movable tables and chairs, which were ordinarily placed in one large circle. The room was also equipped with black shades so that it could be darkened for use of visual aids. A blackboard ran the length of one wall.

The Course

Personal Orientation 1, How to Study, is a two-credit, one quarter course. Classes meet on Mondays, Wednesdays, and Fridays during a ten-week quarter. Administratively it is placed in the General Studies Department of the College of Science, Literature, and the Arts. It is open to students from other colleges, and there are no prerequisites.

The college bulletin describes the course in the following terms:

"Practical assistance to the student in developing efficient methods of study and concentration, organizing material, preparing for examinations, and improving reading ability. Attention to the orientation of students in their attitudes and motivation, and the relation of these to satisfactory performance" (19, p. 74).

Two textbooks were used in the course (1, 17). The topics covered by the texts and by the course include such things as problems of reading, concentration, study methods, use of time, note making, writing, and preparation for examinations.

The Instructors

Both instructors in the study were Ph.D. candidates, with majors in educational psychology and minors in psychology. Both were using the same population and the same experimental procedures with different variables for their doctoral dissertations.

Instructor A had taught Personal Orientation 1 for one quarter prior to the experiment. In addition, he had had two years of previous

teaching and counseling experience at the high-school level.

Instructor B had taught Personal Orientation for five previous quarters. He also had previous teaching experience on the college level.

Both instructors had had an opportunity to try out the different teaching methods before the start of the present experiment. In fact, it was the trial and error and experience with these different teaching methods that lead the instructors to become curious about which method might be superior.

The Instructional Methods

Three different methods of instruction are utilized in the present investigation. These include the instructor-centered method, the student-centered method, and the eclectic method.

Instructor-Centered

The instructor-centered teaching approach follows the line of the traditional lecture method. The philosophy of this method is that the instructor is the authority and his task is to convey information to students so that they may learn and apply.

After all pretest measures had been administered and prior to the beginning of the actual instruction, a syllabus was handed to the students, containing, in addition to general course information, specific reading assignments. The exact number of pages to be read in each textbook prior to each class period was specified in this syllabus.

No attempt was made to encourage student comments or questions in the instructor-centered method. The instructor began each period by calling roll, and then he proceeded to lecture. When, on occasion, a student would raise his hand to ask a question, the instructor would answer it directly and specifically. The instructor did not ask for student opinion or for students to share their experiences on their study problems. When a student's question had been answered, the instructor would continue his lecture or other activity.

The instructor maintained a formal but not a rigid attitude. Humor was occasionally introduced when appropriate. The instructor remained standing at all times.

The students were told to expect unannounced quizzes over reading material. These quizzes were exchanged and scored by students in class. The instructor read the answers, explained why each answer was correct, and answered any remaining questions from students directly.

Students were told that their grade depended upon their test scores from a final examination, a midquarter examination, and the short quizzes.

In addition to lecturing, the instructor occasionally lead students in a period of practice on a certain study technique. All such practice exercises were done at the direction of the instructor at the time and in the manner prescribed by the instructor.

Student-Centered

The student-centered teaching method followed in a general manner the method set forth by Flanders (7). The philosophy underlying this method is that a group of students has within itself the ability to solve its own problems under proper leadership. Since the how-to-study course is organized more to solving problems than toward conveying content, Flanders' technique seems ideally suited for testing the student-centered method.

After the pretests had been administered, the students were handed a syllabus with a statement of course objectives, fundamental considerations, names of the textbooks, and a list of personnel services available at the University. No mention of assignments was made in the syllabus.

Each student was given a name card to place before him on the table, and the informality of the class was emphasized by having each student introduce himself to the class. The instructor lead the class in a discussion of the ways in which a group operates, how it succeeds, and what may cause it to fail. He emphasized that the class belonged to the students, and they were responsible for its operation.

The students were divided into groups of six or less to talk over their study problems and compile a list of problems they felt were most important for them. These problems were listed on the board and organized into topics. The instructor told the class what parts of the textbooks related to each problem they had suggested.

Each student was asked to indicate the topics on which he would be most interested in working. The class was divided into committees on the basis of these choices. The committees were structured by the instructor to a limited extent. They were asked to elect a chairman and a recorder and were held responsible for investigating their topic, reporting their findings to the class, and presenting some class activity in which the entire class would participate.

At this point, the instructor assumed the role of a resource person. He circulated among the committees, answering questions, helping to provide information, and making suggestions when called upon. However, he attempted, as far as possible, to reflect questions in order to place primary responsibility upon the students themselves.

When the planning was completed, each com-

mittee presented its report in turn. The reports were presented in such a manner as to encourage class discussion. During the reports the instructor was seated among the students, and would interject ideas and questions occasionally, much as the students felt free to do.

Activities and reports varied widely. They included such things as panel discussions, open class discussions, short lectures, role-playing, movies, quizzes, reading tests, impromptu themes, surveys, use of the opaque projector, and practice on study skills.

The students were told that their grade would depend upon a final examination, a midquarter examination, any quizzes devised by the committees, and upon the quality of their class participation.

Eclectic

The eclectic method did not confine itself to any one manner of presentation. A combination of many instructional techniques was used. These included many of the techniques used in the two methods previously described.

The class syllabus contained the same information as the student-centered syllabus, plus a listing of the main topics to be covered in the course. The readings relating to each topic were also given.

Name cards and introductions were used to help the group get acquainted and establish a relatively informal atmosphere. A variety of instructional techniques was used with this group. The essential difference between the eclectic and student-centered methods was that in the former the instructor suggested the topics to be covered and the manner in which they were to be presented. The eclectic method differed from the instructor-centered method in that a maximum of student participation was encouraged, although occasional lectures were interspersed with other activities.

The primary technique used was class discussion. Different devices were used to get a maximum of group participation. The buzz session was a technique whereby students would be given a question posed by the instructor and would be asked to discuss it with fellow students sitting near them. This served to warm the students to a general class discussion. The six-by-six technique was a more formal type of buzz session. Groups of six or less students would be formed to discuss a question for six minutes and then report their conclusions back to the class.

The students were told that their grades would depend upon a final examination, a midquarter examination, unannounced quizzes, and the quality of their class participation.

Consistency of Instruction

One of the essential features in the design of this experiment is the principle of replication. To implement this principle it is important that the three treatment methods be consistently carried out as they are intended to be. Three techniques were used to check on the characteristics and consistency of the instruction: coordinated planning by the instructors, a survey administered to the students, and ratings of class recordings.

Coordinated Planning

The two instructors planned in detail the specific content to be covered and the specific instructional techniques to be used in both the instructor-centered and eclectic methods. This was also done for the first part of the student-centered method, when a certain amount of class structuring was necessary.

The coordinated planning was concerned not only with long-range goals but also with the specific daily plans. Before each day's teaching, the instructors prepared their activities jointly. In addition, instructor B, who taught the afternoon sections, routinely checked with instructor A at noon to ascertain whether the instruction had proceeded as planned. When for one reason or another the material covered was not exactly as planned, or instructor A had found it necessary to use some other technique, Instructor B altered his plans so as to conform as nearly as possible to what had occurred in the morning.

After the initial structuring in the student-centered groups, little instructor planning could occur. The groups had been taken over by the students and further instructor planning was not only unnecessary but unwise. However, the instructors did compare notes on the role they assumed in class to keep it as consistently student-centered as possible.

Survey of Opinions—Behavior (SOO-B)

The SOO-B was devised by the instructors to discover whether students perceived the instruction in the way it was intended. Eight items were derived from the definitions of the teaching methods. These are items 11-18 in the Survey of Opinions, a copy of which is found in the Appendix. The first ten items make up SOO-A, which is discussed in Part IV.

Arbitrary weights were assigned to each of the five possible responses of each item. A

weight of four was assigned to a response judged most representative of the instructor-centered method, and a weight of zero to a response most characteristic of the student-centered method. Intermediate weights of three, two, and one were assigned to corresponding intermediate responses. Thus, with these arbitrary weights, the highest possible score, indicating perception of an instructor-centered method, would be thirty-two; the lowest possible score, indicating perception of a student-centered method, would be zero.

The SOO-B was first administered at the end of the third week of instruction to the 166 random and nonrandom students present that day. Using these responses, a new series of weights was devised using the reciprocal averages technique (15). In brief, the method involves a series of successive approximations or iterations to obtain the item weights yielding a maximum internal consistency. Five iterations were necessary to stabilize the item weights for the SOO-B. The reliability of the final iteration was .87.

The SOO-B was administered first at the end of the third week, for the second time at the end of the sixth week, and for the third time during the final week of the quarter. For the first and second administrations all students present, whether random or nonrandom, were included in the analysis. An approximate method of calculating analysis of variance with unequal frequencies in cells was judged to be sufficient for analyzing this data (12, p. 261-266). For the third administration, only randomly placed students were used in the analysis and an exact test of significance was used (4). Only the analysis of the third administration is presented here. The analyses of the first and second administrations yielded approximately the same results. The hypotheses tested in all three administrations of the SOO-B are as follows:

1. No differences exist between mean scores of the SOO-B attributable to different time-instructors.

TABLE 8

MEAN SCORES OF SOO-B FOR EACH SECTION
THIRD ADMINISTRATION
N = 111 Random Students

Instructor	Method			Total
	Student-Centered	Eclectic	Instructor-Centered	
A	7.0952	10.8947	22.0000	13.5667
B	8.4667	9.5203	22.8235	13.6471
Total	7.6667	10.2105	22.8649	13.6036

2. No differences exist among mean scores of the SOO-B attributable to different teaching methods.

3. No interaction effect exists between teaching method and time-instructor.

For the third administration of the SOO-B the students were asked to base their responses on the entire quarter's work. One random student in section 5 missed this administration so the total number of students included is 111. The mean scores are reported in Table 8.

A high score indicates perception of an instructor-centered approach, and a low score indicates perception of a student-centered approach. Inspection of Table 10 reveals that the mean scores are highest for the instructor-centered sections, lowest for the student-centered sections, and intermediate for the eclectic sections. This would tend to indicate that on the average students perceived each class as the instructor intended to teach it. Furthermore, the two instructors were perceived to be operating similarly in each teaching method.

To test the significance of these differences, an exact method of analysis of variance was calculated. The analysis appears in Table 9. This analysis shows that the differences among methods are significant at far less than the one percent level, while differences between time-instructors and the interaction are nonsignificant. The three administrations of SOO-B all

TABLE 9
ANALYSIS OF VARIANCE
SOO-B THIRD ADMINISTRATION
N = 111 Random Students

Source of Variation	df	Sum of Squares	Mean Square	F	Hypothesis Tested
Between Time-Instructors	1	.1057	.1057	<1	Accept
Among Methods	2	4879.8460	2439.9230	164.97*	Reject
Interaction	2	34.1947	17.0973	1.15	Accept
Error	105	1548.3398	14.7461		
Residual	107	1582.5345	14.7900		
Total	110				

* $p < .01$.

yielded the same general results. The students perceived each method as the instructors intended to teach it. Furthermore, the results indicate that the instructors were successful in coordinating their planning so that they consistently portrayed the definition of their roles during the entire quarter.

Class Ratings

As a further check on the consistency of instruction, tape recordings were made of all six sections during one day of instruction. The recordings were analyzed in several different ways. First, ratings were made of the amount of time that the instructor talked in each section. Second, two experienced teachers were asked to listen to each tape and identify which method of instruction was used. And finally, these two raters were asked to write a subjective description of each class after listening to the recordings.

Talk ratio. The tapes for all classes were played while raters tabulated at ten-second intervals whether a student or an instructor was talking. The tabulation was done by alternate ten-second intervals; that is, the first rater tallied at the end of 10, 20, 30, . . . seconds, while the second instructor tallied at the end of 5, 15, 25, . . . seconds. The percentage of total time occupied by the talking of the instructor is indicated in Table 10.

An inspection of this table shows that the highest percentage of instructor talk occurred in the instructor-centered sections and the lowest percentage in the student-centered. In the eclectic sections, both instructors talked about half the time. These results are in agreement with the general principles involved with each type of teaching. Of course it must be remembered that these percentages are based on only one day's instruction and not a necessarily typical day in the judgment of the instructors. They do give a rough idea as to the differences in communication patterns among the three methods.

Identification of methods. The recordings were

TABLE 10
PERCENTAGE OF TOTAL TIME OCCUPIED BY
TALKING OF INSTRUCTOR ON ONE DAY

Instructor	Rater	Method		
		Student-Centered	Eclectic	Instructor-Centered
A	1	6.8	54.5	99.5
	2	4.5	57.6	99.5
B	1	15.3	52.0	84.6
	2	24.0	53.8	90.3

presented to two other experienced teachers in order that they might identify the method used. These two raters were asked to match each tape recording with one of the three methods. They were given the definitions of the three methods, and then listened to the complete tape recording.

At the end of each recording each rater indicated on paper the method he thought had been utilized. Both raters were correct in the identification of all six class periods.

Subjective description. In addition to making an identification of each class, each of the two raters was also asked to write a brief paragraph of description. The descriptions for each class period for each rater are not included here, but the reader may find them in two sources (6, pp. 65-68; 13, pp. 65-68).

Subjective Impressions

Certain unscientific subjective impressions were formed by the instructors during the course of the experiment. Perhaps these impressions could best be expressed as answers to questions that have frequently arisen.

1. *Can one really change his teaching methods as quickly as was necessary to perform this experiment?* Changing one's method of instruction is not as difficult as it may seem. It involves some extra preparation but it can be done. In fact, one of the major subjective outcomes for the instructors was that they discovered they could be more flexible than they had previously thought possible.

2. *How did the students react to the idea of being guinea pigs for the experiment?* Actually, few if any students knew that an experiment on teaching methods was in progress. They were told that the tests they took were part of an evaluation study. This fact was made clear the first day the class met. However, they did not seem to be aware that their section was receiving a different type of instruction than any other How-to-Study section. Apparently, students discussed the activities of the class very little. Only one student mentioned the fact that his syllabus was not like that of another student's from a different class.

3. *Does one have trouble getting students to cooperate in taking the tests?* All testing took place during the class time. The students were promised a copy of their test results at the end of their testing. These were mailed to students who were interested enough to leave a self-addressed, stamped envelope. The students were most cooperative throughout the experiment.

4. *What sort of feelings does one have about each of the methods?* Certainly the instructor-centered method is the most "ego inflating" method from the point of view of the instructor. The students seem to have more "respect" for the instructor. They tend to address him as "Mr." or "Sir," and frequently stop after class to ask the instructor's opinion on a certain study problem. At the end of the quarter there seemed to be more expressions of gratitude from students in the instructor-centered method. In the eclectic and student-centered methods there was less of this particular brand of "respect." Students felt free to disagree with the instructor and less often stopped after class to ask questions. There is no question in the instructors' minds that they actually taught by three different methods.

IV. EVALUATION INSTRUMENTS

The purpose of this part is to describe each of the tests and criterion measures used in the present study. These include the final examination; the *Survey of Study Habits and Attitudes*; the *n* Achievement test; the Opinion, Attitude, and Interest Survey; the Survey of Opinions—Attitudes; the Preferred Instructor Characteristics Scale; the American Council on Education Psychological Examination; and the Personal Data card. The first four of these measures were administered to all students the first week of class before any instruction had been completed. The *first* administration is hereinafter referred to as the *pretest*, while the *second* administration is called the *posttest*.

Final Examination

A 99-item objective final examination was constructed to measure the students' knowledge and understanding of study problems. The process of standardization of the final examination was extended over four quarters of How-to-Study courses offered during 1953-1954 and 1954-1955. During the first three quarters the examination was constantly revised to eliminate the nondiscriminating items. The fourth and final form was administered to a group of 180 students from the fall, 1954. How-to-Study course. Their responses were analyzed in regard to item difficulty, item discrimination, and reliability.

The difficulty of the items was determined by the number of correct responses for each item. There were 46 items in the 25 to 75 per cent difficulty level, 50 above the 75 per cent level, and three below the 25 per cent level. The test was negatively skewed with regard to difficulty, with at least half of the items answered correctly by 75 per cent of the students.

Phi coefficients were calculated by using the upper and lower 27 per cent of the total distribution of scores. Sixty-seven, or two-thirds, of the 99 items correlated .20 or above with the total score. The remaining items fell in the chance area. None were reversal items. By means of Hoyt's (11) analysis of variance technique, the reliability coefficient was found to be .67. This is not as high as is desired for a criterion instrument. The relatively low reliability might have been a function of the restricted range of correct response due to lack of difficulty in the test. However, the reliability was judged sufficiently high to provide a reasonably accurate differentiation between individuals inasmuch as the *F* value "between individuals" in this analysis was significant at the one per cent level. A random sample of fifty students from the current experiment was used to provide a check on the reliability and an almost identical .68 was obtained.

Survey of Study Habits and Attitudes
(SSHA)

The *Survey of Study Habits and Attitudes* (SSHA) was used to measure student changes in study habits and attitudes toward study. Brown and Holtzman (3) have reported most of the essential information about this measure in their manual. The SSHA is empirically keyed for those items which discriminated between students of high and low grade-point average. Because a sex difference was found in the predictive power of the items, the final form was scored on 36 items for men and 29 items for women. In the present investigation, the analysis of results for the SSHA is restricted to males because of the limited number of females.

The internal consistency reliability of the SSHA was estimated by Brown and Holtzman by the split-third method. For men, the Spearman-Brown prophecy formula estimated the reliability to be .92. The test-retest reliability after a two-week interval was .95. On other samples with an eleven-week interval between administrations the estimate was .88.

Estimates of the validity of the SSHA as a predictor of one-semester grade-point averages were obtained from samples of ten different universities across the country. The validity coefficients ranged from .27 to .66. All correlations were found to be positive and significantly different from zero for all schools.

The correlation between SSHA and ACE was found to be consistently lower than the correlations between either of these measures with grades. An exact test of the contribution of the SSHA to the prediction of grade-point average over and above that made by the ACE would utilize the multiple-regression technique. However, this was not reported, so the interpretation that the SSHA contributes unique valid variance to the prediction must be made with caution.

The mean of the total male group of the present investigation was found to be 30.71, which would be equivalent to the 30th percentile on the published norms based on 2,114 men from twelve different colleges. These means are significantly different at the one per cent level. This difference might be expected because many students enroll in the How-to-Study courses because they have academic difficulty and poor study habits. A random sample of 50 students was drawn to determine the internal consistency of the survey for the present investigation. Hoyt's analysis of variance technique gave a reliability coefficient of .88 for this group.

n Achievement (n Ach)

The *n Achievement* test, as its name implies, is a projective device designed to measure one's need for achievement. A complete description of this instrument may be found in a book by McClelland, Atkinson, Clark, and Lowell entitled *The Achievement Motive* (14). In many respects it is very similar to the *Thematic Apperception Test* (TAT) developed by Murray (16), but it differs from it in three major ways. First, although some of the TAT pictures may be used as stimuli, additional pictures have been specially prepared for the *n Ach*. Second, group testing is used. The pictures are flashed on a screen and the subjects write their stories. Third, the scoring system, though derived from some of Murray's concepts, is completely different from that used on the TAT.

The primary assumption in this test is that a person with a need to achieve, a person who might be termed ambitious, driving, or highly motivated, will project this need in the stories he devises. The stories, then, can be objectively scored by counting the number of achievement-related cues. The more highly motivated a person is, the more achievement cues he is expected to project.

Pictures A, B, D, and E (14, p. 375) were administered in that order. One psychologist scored all stories according to the detailed instructions (14, pp. 107-138). His scoring of 80 stories by 20 people correlated .81 with the scoring of these same stories by the authors (14, pp. 346-374). Furthermore, on another

sample of 80 stories scored and rescored after an interval of a month, a score-rescore reliability of .91 was obtained by this psychologist.

A complete and detailed account of the validity of the *n Ach* is given in *The Achievement Motive* (14). No attempt is made here to reproduce this information. However, one study by McClelland (14) conducted on a sample of thirty Wesleyan College male veterans resulted in a correlation between *n Ach* and average grade of .51. When the results of a verbal and mathematical scholastic aptitude test were partialled out, the relationship between *n Ach* and grades was reduced to .39 (which is still significant at the five per cent level of significance). Thus the predictive validity of *n Ach* has been at least partially established by its ability to predict grades.

Opinion, Attitude, and Interest Survey (OAIS)

The OAIS was developed by Benno Fricke (8) to measure certain nonintellectual personality factors important in academic success. In other words, it attempts to measure motivation. This survey consists of 396 statements presented in a paired fashion—a total of 198 items.

The construction of the OAIS was based upon Science, Literature, and the Arts College male students at the University of Minnesota. Only students with ACE percentiles above 54 were included in the criterion groups used for item analysis. Fricke reports about 70 per cent of SLA students score above the 54 percentile on entering University of Minnesota norms (8, p. 164).

On the basis of ACE and honor-point ratio (HPR) Fricke selected a group of students designated as achievers and another group designated as nonachievers. He then performed an item analysis to discover which items of the OAIS (scored configurally) differentiated between these groups. The next step was to establish the usefulness of these items on a cross-validation sample of incoming students. He combined high school rank (HSR), *Ohio State Psychological Examination*, and the OAIS in a regression equation which revealed that the beta weight of the OAIS was significant at the one per cent level of confidence.

The zero order correlation coefficient (*r*) between the OAIS scores and HPR for the total cross-validation sample of 209 students was .40; the *r* for the 138 students having ACE percentiles greater than 54 was .46; and the *r* for the 71 students having ACE percentiles below 55 was .28.

Three limitations should be kept in mind regarding the use of the OAIS in this study:

1. The OAIS was validated on students with

above-average academic ability. The mean ACE (1947 form) raw score was 119, equivalent to the 72nd percentile on norms based on entering freshmen in Minnesota colleges. Students in the present investigation, on the average, scored lower. Their mean ACE (1952 form) raw score was 100, equivalent to the 39th percentile on equivalent norms.

2. The scoring keys were derived from an item analysis of male responses only.

3. The original purpose of the OASIS was to discover personality variables predictive of academic success prior to entrance to college. In the present study this instrument is being applied to students already enrolled in college. Furthermore, these students have undergone a selection process in becoming members of the present study.

Thus both the situation and the nature of the population have been altered in the present application of this instrument.

Survey of Opinions—Attitudes (SOO-A)

The purpose of the SOO-A is to provide a quantitative measure of student attitudes toward class on a favorable-unfavorable continuum. It class on a favorable-unfavorable continuum. It consists of items 1-10 on the Survey of Opinions, a copy of which will be found in the Appendix. Each of these ten statements has five possible responses. The directions instruct the students to answer each question on a percentage basis. That is, he is asked to estimate what percentage of the time he feels a certain way toward the class.

This measure was devised by the instructors in this study simply on the basis of the face validity of each item. Arbitrary weights were assigned to each response to each question. A weight of four was first assigned to the response believed to be most favorable toward the class, three to the next most favorable, on down to zero for the least favorable response to each item. Thus a student with the most favorable attitude toward the class could possibly get a maximum score of forty, and a student with the most unfavorable attitude toward the class could possibly obtain a minimum score of zero.

This questionnaire was administered three times during the quarter to the students in the How-to-Study classes. On the basis of the responses of 145 students taking the second administration, a new series of weights was assigned to the responses using the method of reciprocal averages (15). The weights stabilized after three iterations with few changes in weights and with a reliability of .89. The third administration of the SOO-A is referred to as the SOO-A-3 and was given at the end of the quarter with instructions to consider the entire quarter's work in answering the questions.

The Preferred Instructor Characteristics Scale (PICS)

The primary purpose of this investigation is to discover the effect of different teaching methods on various student outcomes. However, one of the important subsidiary problems is concerned with the effect of students' prejudice for or against a certain kind of teaching method on their outcomes under each condition. There was a need for some instrument to measure this prejudice, and the PICS was designed to meet this need. It is similar in design and theory to an instrument developed by N. L. Gage and others (9, pp. 17-19).

A "cognitive-affective" continuum of instructor characteristics was postulated. A cognitive instructor was defined as one concerned with the intellectual, abstract, subject-matter goals of teaching; the affective instructor was defined as being more concerned with emotional adjustments and student interactions in the classroom. The cognitive instructor corresponds to the instructor-centered instructor of the present investigation, while the affective instructor corresponds to the student-centered type of instructor.

The instructors in this study proposed a number of statements that they believed characterized each of the two type of instructors. Then, to obtain some degree of face validity, they submitted the statements to three advanced graduate students in educational psychology and one instructor in humanities with directions to separate the items according to whether they were "affective" or "cognitive" as defined above.

Only statements which were unanimously classified by all four judges, plus the two instructors, were retained for the scale. Eight cognitive and eight affective items met this criterion.

Each cognitive statement was paired with each affective statement to form a 64-item paired-comparison forced-choice instrument. This form was administered to two fall, 1954, How-to-Study sections to provide the basis for an item analysis. The item analysis revealed that two statements of each type were not discriminating because they were either almost always chosen or almost never chosen. These four statements were discarded. The remaining six cognitive and six affective items were again paired to form a 36-item paired-comparison forced-choice instrument. The order of each type of statement within each item was randomly determined. The resulting form appears in the Appendix.

The PICS is scored in such a way that a high score indicates that the student prefers a cognitive-type approach, while a low score indicates preference for an affective approach. The

maximum score is 36, while the minimum score is zero.

The reliability of the PICS was measured in several ways. It was administered to two different sections of the fall, 1954, How-to-Study class and item analyzed again. Thirty-four of the 36 items yielded a phi coefficient of .20 or higher, indicating that the vast majority of items were successfully discriminating between the upper and lower 27 per cent of the distribution.

A measure of the test-retest reliability was obtained by administering the PICS to 21 night-school students on two occasions with a four-week intervening period. A test-retest reliability coefficient of .88 was obtained.

One further check on the internal consistency of the PICS was made by obtaining a random sample of fifty students from the present investigation. Using Hoyt's analysis of variance technique (11) a reliability coefficient of .90 was obtained. The PICS was administered on the first day of class to students involved in the present study.

American Council on Education Psychological Examination (ACE)

The ACE is widely known and used as a test of scholastic aptitude. Because of this fact a discussion of its construction, administration, scoring, reliability, and validity is unnecessary.

The ACE scores for students in the present investigation were obtained from the files in the Student Counseling Bureau at the University of Minnesota. If a student had not taken the 1952 form of the ACE, his score on one of the other forms was transformed to the 1952 raw score equivalent by means of conversion tables provided by the Student Counseling Bureau. Of the 111 randomly assigned students for whom ACE scores could be located, 58 had taken the 1952 form, 42 had taken the 1947 form, and 11 had taken the 1937 form.

Personal Data Card

A personal data card was employed to collect information about each student's college, class, home and local address, telephone number, age, and sex. In addition, the card was used to record all the test data collected for each student.

Intercorrelation of Instruments

Table 11 presents the intercorrelations of the evaluation instruments based on 80 randomly assigned males on whom complete data were available.

It should be noted that all measures which were administered twice had a test-retest reliability significantly different from zero at the .01 level, yet each of these reliabilities was far from perfect. The two tests which purport to measure motivation (OAIS and *n* Ach) correlate consistently positively with each other, but only on the pretests does this reach significance at the .05 level. Only the final examination shows any significant relationship with the ACE. The two measures of achievement (SSHA and final examination) correlate significantly but not highly with each other.

It should also be noted that the OAIS correlates significantly with all measures of achievement, while *n* Ach reaches significance at the .05 level with only one—the final examination posttest. Liking for class (SOO-A-3) seems to have little relationship with any of the other measures. The PICS is found to correlate significantly with the OAIS on both the pre- and posttest administrations. It also shows a surprisingly high relationship with the SSHA posttest. Perhaps this is an indication that the PICS may have possibilities as a measure of motivation.

V. RESULTS

The primary hypotheses as outlined in

TABLE 11
INTERCORRELATIONS OF THE EVALUATION INSTRUMENTS
(*N* = 80)

Variables	ACE	SSHA (Pre)	SSHA (Post)	Final (Pre)	Final (Post)	PICS	<i>n</i> Ach (Pre)	<i>n</i> Ach (Post)	OAIS (Pre)	OAIS (Post)	SOO-A-3
ACE		.18	-.01	.32†	.27*	-.07	.17	.12	-.02	-.09	-.14
SSHA (Pre)			.69†	.35†	.32†	.12	-.06	.02	.26*	.26*	.07
SSHA (Post)				.32†	.27*	.47†	-.05	.13	.23*	.34†	.26*
Final (Pre)					.59†	.10	.15	.10	.31†	.33†	.20
Final (Post)						.16	.22*	-.03	.25*	.27*	-.11
PICS							.00	.06	.30†	.25*	.06
<i>n</i> Ach (Pre)								.49†	.26*	.20	.00
<i>n</i> Ach (Post)									.18	.17	-.03
OAIS (Pre)										.60	.20
OAIS (Post)											
SOO-A-3											

* Significantly different from zero at the .05 level.

† Significantly different from zero at the .01 level.

Part II are tested first. Following this the subsidiary hypotheses are tested. As specified in Part II, two assumptions for the use of analysis of variance and covariance are tested in these analyses. The variance within each cell is assumed to be homogeneous, and for all covariance problems the linear regression coefficients within each cell are assumed to be homogeneous. In the text of this part, no further mention is made of these assumptions unless they appear to be violated in a particular analysis. Each analysis in the tables of results is marked if either assumption is violated.

For each problem in the investigation an analysis of variance has been calculated on the pretest scores and on the posttest scores. In addition a covariance analysis has been calculated on the posttest scores adjusted for pretest scores. The analysis of most concern is the covariance analysis because any differences in posttest scores are adjusted for differences that might have existed in the initial scores.

Tests of Primary Hypotheses

The primary null hypotheses of the present investigation may be stated as follows:

1. No differences exist among mean scores of the criterion instruments attributable to different methods of teaching.
2. No differences exist between mean scores of the criterion instruments attributable to different time-instructors.
3. No interaction effect exists between teaching method and time-instructor.

The three null hypotheses are tested for each of the four criterion instruments plus the SOO-A-3. For each instrument a separate analysis is made for males and females combined and for males alone. No analysis of females alone is made because of the extremely small frequencies within each cell. On the SSHA no anal-

ysis of males and females combined is made because of the different scoring keys for each sex.

Each analysis of variance (AV) and each analysis of covariance (AC) is summarized in one row of Table 12. The table indicates whether the analysis was performed on pretest scores (pre) or posttest scores (post), and whether males and females combined (MF) or males only (M) were analyzed.

Table 12 reveals that the primary null hypotheses on posttest scores are all accepted on the final examination, the SSHA, and OAIS. However, the analysis of covariance of n Ach posttest scores for males and females combined reveals that the hypothesis of no method difference has been rejected at the .05 level.

What is the nature of the differences between methods? Which method produced the highest level of motivation and which the lowest? The mean scores for each method are presented in Table 13.

The adjusted mean score adjusted for the pretest mean score by means of the regression coefficient. It is the difference in adjusted mean scores that is of importance. The students taught by the eclectic method scored highest. The students from the instructor-centered sections were next highest, and the students in the student-centered sections scored lowest. The chances are less than one in twenty that differences as great as these could have arisen by chance alone.

However, what is the significance of the difference of each method compared with each other? Scheffé (18) has suggested a technique for judging contrasts in analysis of variance. In judging contrasts in analysis of covariance the procedure is virtually identical except that adjusted mean scores are contrasted and the estimated variance of the difference

TABLE 12
SUMMARY OF *F* VALUES TESTING PRIMARY HYPOTHESES

Variable	Total <i>df</i>	Pretest or Posttest	Analysis of Variance or Co- variance	M or MF	<i>F</i> Values for each Source of Variation		
					Among Methods	Between Time- Instruc- tors	Inter- action
Final Examination	111	Pre	AV	MF	<1	<1	2.77
	111	Post	AV	MF	1.02	<1	1.82
	110	Post	AC	MF	<1	1.55	<1
	82	Pre	AV	M	1.24	2.22	2.15
	82	Post	AV	M	<1	<1	3.95
	81	Post	AC	M	<1	<1	1.53
SSHA	82	Pre	AV	M	2.75	<1	1.25
	82	Post	AV	M	1.58	1.71	<1
	81	Post	AC	M	<1	1.55	<1
<i>n</i> Ach	109	Pre	AV	MF	<1	2.20	<1
	109	Post	AV	MF	3.72 ^c	<1	<1
	108	Post	AC	MF	3.41 ^c	<1	<1
	80	Pre	AV	M	1.45	<1	<1
	80	Post	AV	M	1.54	<1	<1
	79	Post	AC	M	<1	<1	<1
OAIS	109	Pre	AV	MF	2.60	<1	<1
	109	Post	AV ^a	MF	1.55	<1	<1
	108	Post	AC ^b	MF	<1	<1	<1
	80	Pre	AV	M	5.06 ^d	<1	1.56
	80	Post	AV ^a	M	1.61	<1	<1
	79	Post	AC ^b	M	<1	<1	<1
SOO-A-3	109	Post	AV	MF	3.69 ^c	<1	6.03 ^d
	80	Post	AV	M	<1	<1	2.86

^a Assumption of homogeneous variances violated.

^b Assumption of homogeneous regression coefficients violated.

^c Significant at the .05 level.

^d Significant at the .01 level.

between adjusted means is slightly increased (5, p. 79).

A null hypothesis of no difference in mean *n* Ach scores for each pair of methods is tested. Scheffé's technique reveals that only the contrast between the eclectic and student-centered methods is significant at the five per cent level. The other contrasts do not reach significance.

Thus we may conclude, with less than one chance in twenty of being wrong, that for this population the eclectic teaching method motivates students to a higher level than does the student-centered teaching method. The instructor-

centered teaching method tends to be intermediate between the other two meth-

TABLE 13
n ACH MEAN SCORES^a ASSOCIATED WITH EACH
METHOD
(Males and Females)

Test	Method		
	Instructor- Centered (<i>N</i> =37)	Eclectic (<i>N</i> =36)	Student- Centered (<i>N</i> =36)
Pretest	15.00	15.54	14.64
Posttest	16.30	17.73	15.03
Adjusted Score	16.32	17.56	15.18

^a These scores include a constant of 10 added to each score to eliminate negative numbers.

ods, but contrasts involving the instructor-centered method do not reach significance.

It is interesting to note that all three null hypotheses are accepted when *males only* are analyzed on the *n* Ach. Apparently, then, the majority of the contribution toward method differences is accounted for by the females.

The present results tend to raise more questions than they answer. In previous research (14) the different motivating conditions were established by means of relatively brief directions. Males were affected by differences in these brief motivating conditions in the expected direction, but females were not. Females seemed to respond in the same manner whether they were motivated under a relaxed condition or under an achievement-oriented condition. The present investigation, by extending the stimulating conditions over a ten-week period of time, may have been long enough to differentially motivate the females. In addition, the social climate may have been sufficiently different under each method to cause different perceptions and reactions. Other explanations are possible. Further research is necessary to isolate the exact nature and etiology of the sex differences associated with *n* Ach.

A significant interaction is evident for the analysis of variance of the SOO-A-3. This interaction precludes any test of the main effects because of the assumption of no interaction in the model used with unequal frequencies in cells (4).

The SOO-A measures attitudes toward class on a favorable-unfavorable continuum. The third administration of the SOO-A is analyzed because it was conducted at the end of all instruction when students were asked to view the entire quarter's work in answering the questionnaire.

TABLE 14
MEAN SOO-A-3 SCORES
(Males and Females)

Time-Instructor	Method			Total
	Instructor-Centered	Eclectic	Student-Centered	
A	28.1	25.8	29.8	28.0
B	32.8	27.5	25.1	28.6
Total	30.3	26.7	27.9	28.3

The reason for the significant interaction can be seen in Table 14. A higher score tends to indicate a more favorable attitude toward class. The interaction is due to the fact that the instructor-centered section of time-instructor B scored higher than the corresponding section of time-instructor A, but the student-centered section of time-instructor A scored higher than the corresponding section of time-instructor B. The differences are not as wide in the eclectic sections. It appears that each instructor (or possibly the time of day) had a method with which he could best please the students, but the same method did not please each group of students equally well for each instructor.

Tests of Subsidiary Hypotheses on Preference Levels

One of the subsidiary problems of this investigation involved the possibility that a preconceived bias concerning the teaching methods might influence the outcomes associated with each method. The PICS, described in Part IV, was designed to measure this bias. On the basis of PICS scores the random students were classified into three categories of approximately equal size. The high scoring third are those students who tend to prefer a more directive or instructor-centered approach; the low scoring third would tend to prefer the student-centered method; while those in the middle might be likely to

prefer some combination of both, presumably the eclectic method.

The exact design for this analysis is set out in Part II. In brief, the purpose of this analysis is to see if differences on the criterion instruments appear to be related to preferences for the different methods as measured by the PICS. At the same time it is possible to discover any interaction effect between preference for instruction and the actual instruction given. A covariance analysis adjusting posttest scores for pretest scores has been carried out to test the following null hypotheses:

1. No differences exist among mean scores of the criterion instruments attributable to different teaching methods.
2. No differences exist among mean scores of

the criterion instruments attributable to different PICS levels.

3. No interaction effect exists between method and PICS level.

The first hypothesis is identical to the first of the primary hypotheses, and consequently the results in the following analysis are identical. It is necessary to include this hypothesis again in order to isolate the variation due to method differences and provide a means for measuring the interaction effect between actual teaching methods and preference for method. The results are summarized in Table 15, which utilizes the same symbols as Table 12.

A highly significant *F* value can be noted for the covariance analysis of SSHA scores. Apparently then, study

TABLE 15
SUMMARY *F* VALUES TESTING SUBSIDIARY HYPOTHESES ON PREFERENCE LEVELS

Variable	Total <i>df</i>	Pretest or Posttest	Analysis of Variance or Co-variance	M or MF	<i>F</i> Values for Each Source of Variation		
					Among Methods	Among PICS Levels	Interaction
Final Examination	111	Pre	AV	MF	<1	1.19	<1
	111	Post	AV	MF	<1	<1	1.76
	110	Post	AC	MF	<1	<1	1.75
	81	Pre	AV	M	<1	1.41	1.58
	81	Post	AV	M	<1	1.82	<1
	80	Post	AC	M	<1	<1	<1
SSHA	82	Pre	AV	M	2.29	<1	<1
	82	Post	AV	M	<1	5.58 ^c	<1
	81	Post	AC	M	<1	7.30 ^e	<1
<i>n</i> Ach	109	Pre	AV	MF	<1	<1	<1
	109	Post	AV	MF	3.86 ^b	<1	1.44
	108	Post	AC	MF	3.41 ^b	<1	1.70
	80	Pre	AV	M	1.80	<1	<1
	80	Post	AV	M	1.40	<1	1.10
	79	Post	AC	M	<1	<1	1.63
OAS	109	Pre	AV	MF	2.51	1.92	<1
	109	Post	AV	MF	1.43	2.48	<1
	108	Post	AC ^a	MF	<1	<1	<1
	80	Pre	AV	M	5.08 ^c	2.59	<1
	80	Post	AV	M	1.38	3.72 ^b	<1
	79	Post	AC ^a	M	<1	1.94	<1

^a Assumption of homogeneous regression coefficients violated.

^b Significant at the .05 level.

^c Significant at the .01 level.

TABLE 16
MEAN SSHA PRE- AND POSTTEST SCORES FOR
THE HIGH, MEDIUM, LOW PICS GROUPS
(Males)

Administration	PICS Levels		
	Low	Medium	High
Pretest SSHA	29.76	29.82	33.08
Posttest SSHA	25.97	29.21	37.38

habit changes do show some relationship to the students' original preferences for a certain type of instruction; and since the interaction is not significant, it can be concluded that this is true regardless of the type of method they actually receive. The nature of this relationship can be seen in Table 16.

Since the differences among methods and the interaction are not significant, the scores from all methods have been pooled to form the mean scores of Table 16. The trend was for the low PICS group to decrease their SSHA scores, for the medium group to rate themselves about the same, and for the high group to increase their SSHA scores over the period of instruction. It appears that the male students who originally preferred a more cognitive or directive teaching method showed the most improvement in their self-ratings on study habits. Those who originally preferred a more affective or student-centered approach tended to rate their own study habits lower after instruction than they did before no matter what type of instruction they actually received.

The reasons for this result are not apparent. However, one hypothesis may be that the PICS discriminates the student who wants to get something out of the how-to-study course from the student who just wants to pick up two credits. On the other hand, since the PICS and the SSHA are positively correlated with each other, this result may simply be an artifact of this relationship, and both tests are simply measures of a positive academic attitude—an attitude that characterizes students who are academically successful.

However, this does not explain why the PICS and the posttest SSHA correlate .47 while the PICS and the pretest SSHA correlate only .12. It is conceivable that the PICS is a measure of motivation toward academic success, but further research is needed to substantiate this hypothesis.

The analysis of variance of OAIS posttest scores for males only does reach significance at the .05 level. The covariance analysis (adjusting for initial pretest differences) is questionable because the assumption of homogeneous regression coefficients has been violated. However, it seems questionable to accept the conclusions of the analysis of variance of unadjusted posttest scores. Differences among PICS levels also appear in the pretest scores, although not quite reaching significance ($F = 2.59$). An exact test is not possible, but it appears that differences among PICS levels on posttest scores are largely the result of initial differences. Furthermore, the initial differences are possibly due to the correlation of the PICS and the OAIS as noted in Part IV.

Tests of Subsidiary Hypotheses on Ability Levels

The second subsidiary problem concerns the scholastic ability level of students and its effect on outcomes under each teaching method. The possibility that bright students might have different outcomes than those less bright, and the possibility that there might be some interaction between ability level and the actual teaching method used are to be tested here. The specific hypotheses to be tested may be stated as follows:

1. No differences exist among mean scores of the criterion instruments attributable to different teaching methods.
2. No differences exist among mean scores of the criterion instruments attributable to different ability levels.
3. No interaction effect exists between teaching method and ability level.

Here again the first hypothesis repeats the first of the primary hypotheses in order to isolate method differences and to provide for the test of interactions. The results are summarized in Table 17.

Examination of Table 17 reveals no significant F values for posttest scores that have not already been mentioned.

TABLE 17
SUMMARY OF *F* VALUES TESTING SUBSIDIARY HYPOTHESES ON ABILITY LEVELS

Variable	Total <i>df</i>	Pretest or Posttest	Analysis of Variance or Co- variance	M or MF	<i>F</i> Values for Each Source of Variation		
					Among Methods	Among ACE Levels	Inter- action
Final Examination	111	Pre	AV	MF	< 1	1.53	< 1
	111	Post	AV	MF			< 1
	110	Post	AC	MF	1.05	2.73	< 1
	81	Pre	AV	M	< 1	2.11	< 1
	81	Post	AV	M	1.37	3.40 ^b	< 1
	80	Post	AC	M	< 1	2.30	< 1
SSHA							1.05
	81	Pre	AV	M	2.59	< 1	< 1
	80	Post	AV	M	1.13	< 1	< 1
<i>n</i> Ach		Post	AC	M	< 1	1.17	< 1
	108	Pre	AV	MF	< 1	2.39	< 1
	108	Post	AV	MF			
	107	Post	AC	MF	4.04 ^b	< 1	1.43
	79	Pre	AV	M	3.51 ^b	< 1	1.97
	79	Post	AV	M	1.80	1.77	< 1
	78	Post	AC ^a	M	1.66	< 1	< 1
OAIS					< 1	< 1	1.29
	108	Pre	AV	MF	2.53	< 1	< 1
	108	Post	AV	MF			
	107	Post	AC	MF	1.45	1.32	1.35
	79	Pre	AV	M	< 1	< 1	1.02
	79	Post	AV	M	4.88 ^b	< 1	< 1
	78	Post	AC	M	1.60	< 1	< 1
					< 1	< 1	1.02

^a Assumption of homogeneous variances violated.

^b Significant at the .05 level.

Apparently, scholastic aptitude as measured by the ACE bears no significant relationship to outcomes on the variables used in this study. There does not appear to be any tendency for a bright student to be more successful under one teaching method than another.

VI. SUMMARY AND CONCLUSIONS

The primary problem in the present investigation is to discover any achievement or motivational differences that may result from three different methods of teaching in a how-to-study course. The population consists of all University of Minnesota students who elected to take Personal Orientation 1, How to Study, in the winter quarter of 1955. Approximately 63 per cent of the 192 students

enrolled were randomly assigned to three teaching methods taught by two instructors at two different times of the day. Although nonrandom students do not appear to be essentially different from the random students on any of the characteristics examined, the analysis of results is confined to the random students.

The three teaching methods are termed instructor-centered, student-centered, and eclectic. The instructor-centered method emphasizes the intellectual content of the course and consists primarily of lectures and instructor-directed activity. The student-centered approach tends to emphasize the more affective aspects of the classroom and deals with student problems by means of committee work and student-led discussions. The

eclectic method consists of a combination of the previous emphases and proceeds primarily by means of instructor-led class discussion interspersed with a variety of other techniques. Checks on the consistency of instruction by each instructor in each method by means of coordinated planning, a student rating on instructor behavior, and judges' ratings of class recordings tended to affirm the consistency of instruction.

A design embodying the principles of replication, randomization, and control is utilized. Null hypotheses are constructed to test the relationship of method, time-instructor, and interaction to pretest and posttest measures of motivation and achievement. Subsidiary analyses of the effect of original preference for instructional method and ability-level on motivational and achievement outcomes associated with each method are conducted. Analysis of variance and covariance by least-squares analysis is the primary statistical tool.

Five criterion instruments are utilized. The final examination is a 99-item, objectively scored, test of knowledge about the course content. *The Survey of Study Habits and Attitudes* (SSHA) was used to obtain student self-ratings of study habits and attitudes. The *n* Achievement test (*n* Ach) is a projective technique designed to measure achievement motivation. The Opinion Attitude and Interest Survey (O AIS) is a configurally scored inventory which has been shown to contribute significantly to the prediction of honor-point ratio. All four of these instruments were administered both before and after the instructional period. The fifth instrument is the Survey of Opinions—Attitudes (SOO-A), a questionnaire designed to measure student attitudes toward class on a favorable-unfavorable continuum.

The results of this study may be stated as follows:

1. Significantly different motivational outcomes as measured by the *n* Ach are found among the three teaching methods for males and females. Students in the eclectic section were most highly motivated as measured by *n* Ach; the instructor-centered students were second; and the student-centered students showed the least increase. These method differences disappear when males only are analyzed, indicating the major contribution of females to these differences.

2. Differences in outcomes as measured by the final examination, the SSHA, and the O AIS are not significantly different for the three methods.

3. On the final administration of the SOO-A for males and females an interaction of method with time-instructor is found. The method which was most successful in pleasing the students under one instructor (or possibly at one time of day) was not most successful under another instructor (or at another time of day).

4. When students are categorized on the basis of their original preference for teaching method, it is found that students who originally expressed a preference for a more cognitive-type instruction increased their self-ratings of study habits and attitudes (SSHA). Students who originally preferred a more affective or student-centered type of instruction tended to lower their self-ratings. This was true regardless of the type of instruction they actually received. On the other variables no such differences were observed.

5. When students are categorized on the basis of their scholastic-ability level as measured by the ACE, no significant motivational or achievement outcomes are found in relation to ability level. Fur-

thermore, there is no tendency for bright students to have any different outcomes under one teaching method than another.

Previous research on teaching methods has dealt with extreme teaching methods. This may have been done to maximize the possibility of obtaining significant differences. The present investigation illustrates the importance of considering an intermediate *eclectic* method combining the advantages of both the typical lecture and the typical group approach. Many of the studies on teaching methods cited by Farquhar (6) could be profitably repeated with the introduction of an eclectic method. It could well be that a combination method is far superior to either extreme in implementing many

educational objectives. Certainly the results of this study point in that direction.

In designing an investigation similar to the present one several improvements could be made. Many students were lost to the analysis because the randomization procedure was incomplete. If it were possible administratively to assign all students randomly, the precision of the experiment would be increased. Another administrative problem concerns the use of a no-treatment control group. No evidence can be obtained from the present design as to whether instruction by any method is more effective than no instruction at all. Where possible, future research designs of this nature could be improved by more complete randomization and the institution of a no-treatment control group. Complete details of this investigation, including reviews of literature and the raw data, may be found in two sources (6, 13).

Further research in other subject matter areas and with other populations is necessary to substantiate the results of this study. The study poses many questions which can only be answered by controlled experimentation.

APPENDIX

PREFERRED INSTRUCTOR CHARACTERISTICS SCALE

Directions:

What kind of an instructor do you prefer? In the following items you will find two instructor characteristics paired. From each pair choose the one characteristic you most prefer. Then mark your choice in the proper column on the special answer sheet. Do not omit any items. This is to find out your preferences. *There are no right or wrong answers.*

I prefer an instructor who:

1. *a.* is an expert.
b. treats us as mature people.
2. *a.* makes the classroom pleasant.
b. thinks logically.
3. *a.* understands our point of view.
b. is well known in his field.
4. *a.* is dedicated to his students.
b. is dedicated to his subject.
5. *a.* thinks logically.
b. is friendly.
6. *a.* is well known in his field.
b. makes the classroom pleasant.
7. *a.* is interested in us.
b. covers all the material.
8. *a.* is dedicated to his students
b. knows the theoretical background of his subject.
9. *a.* thinks logically.
b. treats us as mature people.
10. *a.* is friendly.

- b.* is well known in his field.
11. *a.* covers all the material.
b. understands our point of view.
12. *a.* is interested in us.
b. is dedicated to his subject.
13. *a.* is an expert.
b. is dedicated to his students.
14. *a.* is well known in his field.
b. treats us as mature people.
15. *a.* covers all the material.
b. makes the classroom pleasant.
16. *a.* understands our point of view.
b. is dedicated to his subject.
17. *a.* is interested in us.
b. knows the theoretical background of his subject.
18. *a.* is friendly.
b. covers all the material.
19. *a.* makes the classroom pleasant.
b. is dedicated to his subject.
20. *a.* knows the theoretical background of his subject.
b. understands our point of view.
21. *a.* is interested in us.
b. is an expert.
22. *a.* is dedicated to his students.
b. thinks logically.
23. *a.* treats us as mature people.
b. covers all the material.
24. *a.* is dedicated to his subject.
b. is friendly.

25. a. makes the classroom pleasant.
b. knows the theoretical background of his subject.
26. a. is an expert.
b. understands our point of view.
27. a. is dedicated to his students.
b. is well known in his field.
28. a. is dedicated to his subject.
b. treats us as mature people.
29. a. is friendly.
b. knows the theoretical background of his subject.
30. a. is an expert.
b. makes the classroom pleasant.

31. a. thinks logically.
b. is interested in us.
 32. a. treats us as mature people.
b. knows the theoretical background of his subject.
 33. a. is an expert.
b. is friendly.
 34. a. thinks logically.
b. understands our point of view.
 35. a. is interested in us.
b. is well known in his field.
 36. a. is dedicated to his students.
b. covers all the material.
- Check to see if you left any blanks.*

REFERENCES

1. BIRD, C., & BIRD, DOROTHY M. *Learning more by effective study*. New York: D. Appleton-Century, 1945.
2. BIRNEY, R., & MCKEACHIE, W. The teaching of psychology: a survey of research since 1942. *Psychol. Bull.*, 1955, 52, 51-68.
3. BROWN, W. F., & HOLTZMAN, W. H. *Survey of study habits and attitudes*. New York: Psychological Corp., 1953.
4. BUCHMAN, R. The least squares analysis of a 2-way factorial experiment with unequal frequencies in the cells. Mimeographed paper, Bureau of Educational Research, Univer. of Minnesota, 1954.
5. COCHRAN, W. G., & COX, GERTRUDE M. *Experimental designs*. New York: Wiley, 1950.
6. FARQUHAR, W. W. An investigation of the relationship of three teaching methods to student behavior in a how to study course. Unpublished Ph.D. thesis, Univer. of Minnesota, 1955.
7. FLANDERS, N. A. *Teaching with groups*. Minneapolis: Burgess Publishing Co., 1954.
8. FRICKE, B. G. The development of an empirically validated personality test employing configural analysis for the prediction of academic achievement. Unpublished Ph.D. thesis, Univer. of Minnesota, 1954.
9. GAGE, N. L., LEAVITT, G. S., & STONE, G. C. Teachers' understanding of their pupils and pupils' ratings of their teachers. *Psychol. Monogr.*, 1955, 69, No. 21 (Whole No. 406).
10. GOULDEN, C. H. *Methods of statistical analysis*. New York: Wiley, 1939.
11. HOYT, C. J. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
12. JOHNSON, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
13. KRUMBOLTZ, J. D. An investigation of the effect of three teaching methods on motivational outcomes in a how to study course. Unpublished Ph.D. thesis, Univer. of Minnesota, 1955.
14. MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
15. MITZEL, H. E., & HOYT, C. J. A methodological study of reciprocal averages technique applied to an attitude scale. *J. counsel. Psychol.*, 1954, 1, 256-259.
16. MURRAY, H. A. (Ed.) *Explorations in personality*. New York: Oxford Univer. Press, 1938.
17. ROBINSON, F. P. *Effective study*. New York: Harper, 1946.
18. SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, 40, 87-104.
19. *Bulletin of the University of Minnesota, College of Science, Literature and Arts*, 1953, 56, No. 23.

(Accepted for publication March 7, 1957)

SURVEY OF OPINIONS

Directions:

The purpose of this survey is to determine your opinions about this course. Please be frank and honest in your answers. Naturally, this will not affect your grade in any way. It is important for the instructor to know how the class feels on each of these questions.

To help you in answering, the following terms have been defined on a percentage basis as follows:

A—Almost always—from 86 to 100 per cent of the time

G—Generally—from 66 to 85 per cent of the time

F—Frequently—from 36 to 65 per cent of the time

S—Sometimes—from 16 to 35 per cent of the time

R—Rarely—from 0 to 15 per cent of the time

Circle the letter corresponding to your own opinion:

- A G F S R 1. I enjoy this class.
 A G F S R 2. I feel this class has been valuable to me.
 A G F S R 3. I feel "at home" in this class.
 A G F S R 4. This class has been well taught.

- A G F S R 5. I am glad I took this course.
 A G F S R 6. I think class time is well spent.
 A G F S R 7. I look forward to each class meeting.
 A G F S R 8. This class is interesting to me.
 A G F S R 9. I feel the instructor wants me to do well.
 A G F S R 10. I feel this is an important class for me.
 A G F S R 11. The instructor encourages us to answer our own questions.
 A G F S R 12. The problems we work with arise out of the questions we suggest.
 A G F S R 13. We have opportunities to compare our ideas and methods with others in the class.
 A G F S R 14. We evaluate our own activities in class.
 A G F S R 15. We work with the rest of the class to get answers to our questions.
 A G F S R 16. The instructor lectures.
 A G F S R 17. The instructor determines what topics will be discussed.
 A G F S R 18. The assignments are very clear and definite.

SCORING KEYS

Preferred Instructor Characteristics Scale (PICS)

A plus one is given for each of the following cognitive responses:

1. a 7. b 13. a 19. b 25. b 31. a
 2. b 8. b 14. a 20. a 26. a 32. b
 3. b 9. a 15. a 21. b 27. b 33. a
 4. b 10. b 16. b 22. b 28. a 34. a
 5. a 11. a 17. b 23. b 29. b 35. b
 6. a 12. b 18. b 24. a 30. a 36. b

Survey of Opinions

The following weights, derived from the final iteration, were used:

SOO-A						SOO-B					
Item	Response					Item	Response				
	A	G	F	S	R		A	G	F	S	R
1.	4	3	1	0	0	11.	1	1	3	4	4
2.	4	3	2	1	1	12.	0	2	3	4	4
3.	4	2	3	1	0	13.	0	2	2	3	4
4.	4	3	2	0	0	14.	0	2	2	3	4
5.	4	3	2	1	0	15.	0	1	2	4	4
6.	4	3	2	1	0	16.	3	3	1	1	0
7.	4	4	2	1	1	17.	3	2	2	1	0
8.	4	3	2	1	0	18.	3	1	0	0	1
9.	3	3	2	2	0						
10.	4	3	2	1	0						

A Criterion For Counseling¹

CLIFFORD P. FROEHLICH

University of California, Berkeley

INTRODUCTION

ONE of the most obvious objectives of counseling is to increase the counselee's knowledge of himself. This objective is justified on the grounds that a client must know himself if he is to make satisfactory adjustments to the difficulties which brought him to counseling. This justification is particularly pertinent when the problems of educational and vocational planning are considered. A client must have a reasonably accurate estimate of his mental ability *before* he can intelligently decide whether to enter college or to pursue a professional career. Likewise, he needs some knowledge of his interests *before* he can logically choose a curriculum or a vocation in harmony with them. The word "before" has been italicized in the preceding sentences to highlight a significant point in connection with this investigation which studies a criterion of counseling effectiveness. The criterion compares one perception of self with an external measure of self, and notes agreement prior to counseling. A second comparison of self-perception and external measure is made after coun-

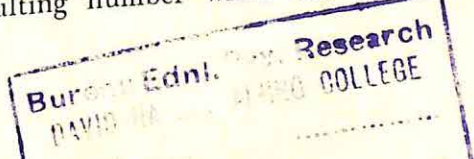
seling, and agreement again is noted. The criterion is the change in agreement between the first and second comparison. The criterion studied combines good features of both sociological and psychological criteria. The sociological aspects of the criterion are, of course, found in the postcounseling behavior of the client. He cannot score favorably on the criterion, except by accident, unless he has increased self-knowledge. Hence, an improvement in such knowledge creates a "before" situation, subsequent to which the client is more able to show improved behavior.

The criterion in this investigation might be viewed also as a psychological one. It differs, however, from the Q sort and similar criteria used by Rogers and others in that it is based on an objective measure, a test score. Rogers (8) simply compared one perception of self with another perception of self in order to note changes. The present study compares self-perceptions obtained in the form of self-ratings with objective test data.

Terms Used in this Study

For the sake of brevity, certain terms will be used to denote concepts which would require fuller explanation if the following definitions were not made. The terms "congruence" and "agreement" are used interchangeably. They are used to indicate coincidence between two or more measures. "Score" represents the resulting number when the individual's

¹This study was aided by a grant from the Research Fund of the Department of Education, University of California. The cooperation of Robert Brownlee, Principal, Demonstration Secondary School, made it possible to conduct the study in his school. I. Aileen Poole assisted in the collection of the data. E. Wayne Wright and Joseph Martin helped in the analysis of the data, with the use of the Computer Laboratory of the University.



actual test score is converted (in the manner to be described) to its corresponding step on a five-point scale. "Score" means, then, the quintile rank-equivalent of the test score. The word "rating" is used to indicate the numerical step on a five-point scale which the individual checked when asked to rate himself.

DESIGN OF STUDY AND COLLECTION OF DATA

The purpose of this study was to investigate a variable which appears to possess the quality of being a sociological-psychological criterion. To accomplish this purpose, specific questions were formulated to be answered by the data collected in the experimental situation.

1. The first of these questions was: *Is there a significant difference between the agreement of precounseling rating with score and the agreement of postcounseling rating with score?* A related question was also asked. *Are the counselees' changes in agreement in the direction of greater agreement?* In other words, do the comparisons made before counseling differ significantly from those made after counseling; and does counseling increase the congruence of self-ratings and scores? It was anticipated that if these questions were answered in the affirmative, the following one could be asked of the data.

2. *Does the counseled group at the conclusion of the experiment show more ratings which are congruent with scores than the noncounseled group?* This question formalized the expected differences between counseled and noncounseled groups of students. In contrast, the first question made explicit the idea that the number of agreements would change from before counseling to after counseling. In this case, each counselee acted as his own control. In the other, comparisons were made between groups of stu-

dents unlike in their participation in counseling.

Locale and Population of the Study

The basic data for this study were collected during the summer of 1953 at the Demonstration Secondary School operated by the University of California in cooperation with the Oakland, California, public schools. The summer high school completed its sixteenth session in 1953, at which time the total enrollment was 951 students. Sixty-nine per cent of the students were residents of Oakland, 30 per cent came from other California communities, and 1 per cent from outside of California. The students came from 58 high schools. Their median age was 14 years, 11 months, and their median grade placement during the preceding semester was the second semester of grade 10.

A poll of the students' reasons for attending the summer school indicates that 27 per cent were there to repeat a subject or to improve the mark, 25 per cent to gratify special interests, and the remainder were completing requirements for college entrance or high school graduation. In comparison with national norms, the student population was skewed in the direction of superior ability on a test of mental ability.

All students who during the preceding semester had been enrolled in grades nine through twelve were required to take a battery of tests during their so-called "guidance period." This period was regularly scheduled but carried no credit. The students, after testing was completed, were not expected to be at school during their guidance period.

The purpose of the testing program was explained to the students at the first testing session. They were told that they were participating in an experimental testing program which was conducted for research purposes by the University of California. They were informed also that the test results would be available to them if they wished to talk with a counselor. No effort was made to induce students to make use of the counselors. At the first testing session, students were asked whether or not they wanted to talk with a counselor. The data in Table 1 indicate the number who wished to be counseled as well as the number who were actually counseled. The discrepancy between the total of 951 students enrolled in the school and the 485 in Table 1 is accounted for by a variety of factors. Some were excused from the testing program if their out-of-school employment made it necessary, as were also foreign students whose knowledge of English was not sufficient to complete the tests. Approximately 200 students were eliminated from the study because they had not been enrolled in a high school during the preceding semester.

TABLE 1
STUDENTS INCLUDED IN THE STUDY

Do you wish to be counseled?	Male Students		Female Students		Total	
	Counseled	Not Counseled	Counseled	Not Counseled	Counseled	Not Counseled
Yes	113	63	107	41	220	104
No	9	63	8	36	17	99
Did not answer question	3	21	5	16	8	37

Others were eliminated because their criterion or test data were incomplete. In all, 466 of the 951 students were not used in the present investigation.

The Test Instruments

The following tests were used: *SRA Primary Mental Abilities*; *Intermediate (9)*; *Kuder Preference Record-Vocational, Form C (6)*; *Test of Mechanical Comprehension, Form BB (1)*; and *SRA Youth Inventory, Form A (7)*.

The tests of interest, mental ability, and mechanical comprehension were selected because they were judged to be typical of those used in high school counseling programs. Studies of their reliability and validity have been reported in the journal literature and publishers' manuals. For each of these tests a listing of references, as well as several critical reviews, are included in the three most recent editions of *Buros' measurement yearbooks* (2, 3, 4). All appear to be sufficiently reliable for the purposes of this investigation.

The validity of the tests used in this investigation deserves a general comment. Although they are generally accepted as being valid instruments, in one view, their validity may be considered as unrelated to the study. In this instance the study is viewed as dealing with the question of whether counseling increases the congruence between the counselee's self-rating and his score. In other words, does counseling teach a person his score so that he does reproduce it in the form of a self-rating. It is not a question of whether the score on a test of, say, outdoor interests is a valid measure of outdoor interests. The fact that the counselor accepts the score as an indication of outdoor interests and interprets it as such to the counselee lends credence to the idea that if counseling is effective counselees should show more agreement between their self-ratings and scores on tests of outdoor interests after counseling than they did before.

The criterion used in this study is concerned with the observed change when the agreement

of precounseling score and ratings is compared with the agreement of postcounseling scores and ratings. Essentially the criterion is anchored to agreement of score with rating, rather than agreement of the rating with some absolute or perfect measure of the characteristic upon which the counselee rates himself. If in counseling the score is interpreted as being indicative of a more perfect measure of the characteristic (whether or not it actually is beside the point) then the criterion should be influenced in the hypothesized manner implied in the two major questions stated previously.

The validity of the *SRA Youth Inventory* is not germane since this check list was not used as one of the variables in the criterion. It was employed, rather, in an attempt to identify differences between persons who ranked high and those who ranked low on the criterion. In a sense, an attempt was made to validate it against the criterion.

All the raw scores on the tests were converted to quintiles on the basis of the published norms with the exception of the mechanical ability test. Here no appropriate norms were available, hence quintiles were based on a sample of students making up the population used in this study.

The Ratings

The self-ratings were secured by means of a scale designed for the study. Students were asked to rate themselves on a five-step scale for each of the ten interest areas measured by the *Kuder Preference Record*, and also their mental and mechanical ability. The interest or ability to be rated was described in nontechnical language.

Students were asked to complete the rating scale at the beginning of the first testing session. In the last week of the summer term, seven weeks after students completed the first rating scale, a sec-

ond one was administered. It was identical with the first except that the instructions were changed to include a rationale for the request for repeat ratings. The second ratings were made during regular class periods.

The criterion investigated in this study was based upon these self-ratings. Such ratings have been utilized in a variety of research studies and have themselves, in some instances, been subjected to study. The literature pertinent to the topic "ratings" is voluminous; that pertaining to self-ratings is more limited. It was recognized that self-ratings had a limitation in terms of reliability. Other measures might have greater reliability but none seemed to hold as much promise for studying changes in self concept as did self-ratings. Indeed, the task of rating oneself requires one to reveal aspects of one's self concept or to resort to some sort of deception in order to conceal it. This is a compelling reason for using changes in self-estimates as a criterion of counseling. But an even more pertinent reason is to be found in the very undependability of the ratings. One of the objectives of counseling is to eliminate or at least reduce the influence of factors which contribute to the unreliability of ratings. Counseling should improve the accuracy of the counselee's self concept. If it does, it would make it possible for him to rate himself more accurately.

At the outset of the study another limitation of the criterion was recognized: there is more involved in self-rating than just knowledge of

self. A counselee may, for example, "know" on the intellectual level that his performance on a test of mechanical comprehension is equivalent to the second centile of entering students in an engineering school. But, if he dreams of bridging the Bosphorus, irrigating the Sahara, or inventing a space ship, he may not really believe his mechanical comprehension is limited. His concept of himself as a successful engineer, figuratively speaking, insulates him from the threat of accepting the meaning of his low mechanical comprehension test score. The net result is that he rates his mechanical comprehension higher than his score rates him. The implication of this limitation for this study is that the criterion is not a measurement of changes of actual self concept but rather is assumed to be a reflection of those changes.

The reliability of the ratings used in this study was estimated from data collected in the same school situation but in a different summer. Students were asked to rate themselves and then a week later to repeat the process. No tests were given, nor were interviews held during the interval between ratings. The reliability coefficients reported in Table 2 were computed by using the customary formula for product-moment correlation and were corrected for coarse grouping in the manner described by Wert, Neidt, and Ahmann (10). All of the uncorrected coefficients are significant at the .05 level except that for "persuasive." Of the 13 corrected coefficients, 9 are significant at the .01 level. These coefficients seem to indicate that the ratings used in this study have a degree of reliability not unlike that reported in other studies of ratings. True, they

TABLE 2
RELIABILITY COEFFICIENTS OF RATINGS

Area	N	Correlation of Ratings			
		Uncorrected		Corrected	
		Reliability Coefficient	Level of Significance	Reliability Coefficient	Level of Significance
Interests:					
Outdoor	31	.40	.05		
Mechanical	31	.88	.01	.43	.05
Computational	31	.62	.01	.90	.01
Science	31	.39	.01	.66	.01
Persuasive	30	.25	.05	.42	.05
Artistic	31	.43	.18	.27	.16
Literary	30	.74	.05	.46	.01
Musical	31	.64	.01	.77	.01
Social Service	31	.51	.01	.68	.01
Clerical	31	.38	.05	.51	.01
				.41	.05
Abilities:					
Mental	27	.78	.01		
Mechanical	30	.46	.01	.81	.01
				.50	.01

are not as high as might be desired. Nevertheless, as the writer stated earlier, ratings appear to him to have a potential for assessing self concept and this was an important factor in choosing them.

The Counselors

The persons who did the counseling in the Demonstration Secondary School were all graduate students specializing in student personnel and counseling psychology in the Department of Education, University of California. At the time of their participation in this study, the counselors were enrolled in the university summer session working toward advanced degrees in counseling psychology. During the preceding semester nine had been employed as secondary school teachers, six as counselors at the college level, four as teacher-counselors in high schools, three as full-time counselors at the secondary level, two as school social workers, one as teacher-psychologist, one as head counselor, and one as guidance supervisor.

The counselors chose their own orientation to counseling. None was completely nondirective. All were aware of the fact that the pupils had been told first of the availability of counseling when they reported to the testing sessions. The promised opportunity to learn and talk over the test results was used undoubtedly by the administrators as a motivating factor during the series of five testing sessions. Hence, some pupils who reported to the counseling office came for the express purpose of test interpretation. Counselors made an effort to identify pupils who had an inadequate notion of the kind of relationship that the counselor was willing to establish. They made a special effort to make certain that all counselees realized that the counselor was willing to talk about things other than test scores.

The Criterion

In counseling practice, as has been pointed out, one of the major objectives is to help the client improve his knowledge of himself. The assumption is made that for a client to make wise choices which involve him, it is necessary for

him to know himself. In some counselors' orientation, the client can best discover self-knowledge in a nondirective or client-centered relationship. Other counselors consider it a more meaningful experience if the counselor plays a more active role. Although counselors vary greatly in their relative adherence to either of these two points of view, there is rather general agreement among them that an important outcome of counseling is an increase in the client's knowledge of himself. The criterion investigated in this study was designed to measure this outcome.

The criterion was change in agreement between a person's rating of his ability or interest and an objective measurement thereof. In short, the client's ratings on a five-point scale in each area were compared with quintile ranks based on his test scores in corresponding areas. In terms of the desired outcome—improvement in knowledge of self—certain assumptions had to be made concerning the criterion data.

Although it was recognized that the test scores were fallible and perhaps might not be accurate, it was assumed that they did reflect an image of the client. From this assumption it followed that if the client had accurate knowledge of self his view of himself would correspond with the test-revealed one. Conversely, it was expected, therefore, that clients would learn about themselves from test scores.

A second assumption was made that if a person knew what his tested abilities and interests were he would be able to rate them accurately. Conversely, if he did not know what his abilities were, then he would be unable to rate them accurately.

The third assumption was made that although the subjects of this study would undoubtedly have a considerable amount of information about themselves, the process of test-taking and counseling would improve this knowledge. These three assumptions were basic to the design of the study.

The data are presented for three groups of students: *Counseled, counseled-without-test, and not counseled.*

In the *counseled* group were included those students who were counseled and with whom the counselor discussed the results of the test. It should be noted that the counseled group varied from test to test; that is, not all persons who were counseled had all test results interpreted for them. One counselee, for example, may have had the interest inventory and mental ability test interpreted while another may have discussed the interest inventory and the mechanical aptitude test with the counselor. Both counselees would be included in the counseled group for the ten scores of the interest inventory. The first counselee would be included in the counseled group for the mental ability test and in counseled-without-test group for the mechanical aptitude test. Conversely, the second would be included in the counseled group of the mechanical aptitude test and counseled-without-test group on the mental ability test.

The second group, *counseled-without-test*, includes those persons who were counseled but for whom the particular test was not interpreted. Noninterpretation of tests was occasioned by several factors such as a delay in scoring the test so that it was not available at the time of counseling or the judgment of the counselor that the test score would not make a significant contribution to the counseling.

The third group, "*not counseled*," includes those individuals who took the tests, but who were not counseled. Counseling was available for all who desired it but no effort was made to induce students to seek counseling. Of the 116 students who indicated at the first testing session that they did not wish to be counseled, only 17 later changed their minds and sought counseling.

FINDINGS

The data concerning each test were separated into three groups according to whether or not the individual was counseled, counseled-without-test, or not counseled. In order to study the agreement between scores and ratings, each individual's score and corresponding rating were compared. If both fell on the same numerical step of the five-point scale, it was tallied as agreement. That is, if the individual's score was a "five" and his rating a "five," the values were considered in agreement. But if his score was "five" and his rating "four," the values were tallied as disagreement. The agree-

ment between first rating and score was determined, as was also the agreement between second rating and score. Each individual on the basis of agreement between score and ratings of each area was classified into one of four categories, namely:

Category 1. The score and first rating disagreed and so did the score and second rating. Persons in this category consistently rated themselves inaccurately.

Category 2. The score and first rating agreed and so did the score and second rating. These individuals were consistent in rating themselves accurately.

Category 3. The score and first rating disagreed, but the score and the second rating agreed. Included in this group are those individuals whose first rating was incorrect but who correctly rated themselves on the second rating. They changed from "disagree" to "agree."

Category 4. The score and first rating agreed but the score and second rating did not. These persons changed their rating from an accurate one to an inaccurate one. They went from "agree" to "disagree."

Percentages based on the data tabulated according to the above-described categories are shown in Table 3. It is obvious that the majority of individuals' ratings consistently disagreed with their score. This does not mean that they did not change their ratings. They might have rated themselves too low on the first rating and too high on the second; in both instances score and rating disagreed. The median percentage for this category for the counseled group was 55, for the counseled-without-test group was 54, and for the not-counseled group was 61.

These median percentages are indicative of a situation of vital concern to counselors. Why is it that the majority of pupils in this study were unable to rate accurately their abilities on even such a gross scale as the five-point one used here? In a preceding section it was pointed out

TABLE 3
PERCENTAGES FOR CATEGORIES OF AGREEMENT BY GROUPS AND AREAS

Area	Counseled					Counseled-Without-Test					Not-Counseled				
	N	Consistently Disagree	Consistently Agree	Disagree to Agree	Agree to Disagree	N	Consistently Disagree	Consistently Agree	Disagree to Agree	Agree to Disagree	N	Consistently Disagree	Consistently Agree	Disagree to Agree	Agree to Disagree
Interests:															
Outdoor	175	69	12	13	06	66	64	11	11	14	229	72	08	10	10
Mechanical	174	55	17	13	15	66	64	20	08	08	230	70	11	11	08
Computational	175	50	15	17	18	66	43	26	17	14	229	60	14	14	16
Science	175	47	20	19	14	66	57	15	11	17	230	56	18	11	15
Persuasive	175	53	18	20	09	66	54	14	18	14	229	63	12	13	12
Artistic	175	58	20	11	11	66	46	17	20	17	230	47	19	16	18
Literary	175	52	22	15	11	66	54	23	14	09	229	59	14	14	13
Musical	175	53	21	15	11	66	47	24	15	14	230	51	26	12	11
Social Service	175	55	17	19	09	66	45	18	17	20	229	63	11	12	14
Clerical	175	56	17	12	15	66	48	14	23	15	230	61	16	12	11
Abilities:															
Mental	150	68	15	11	06	75	66	17	09	08	218	66	11	10	13
Mechanical	79	60	14	15	11	158	62	13	10	15	229	61	13	14	12
Musical	99	57	19	10	14	128	67	16	09	08	212	64	16	08	12

that self-ratings were used because they required the subject to reveal aspects of his self concept or resort to deception. The data in this study cannot yield the reason for the inaccuracies. They do, however, hint at the possibility that lack of agreement between rating and test score was brought about not so much by a lack of "knowledge" of self as it was by a lack of "acceptance" of the knowledge or an unwillingness to reveal the knowledge through ratings. Consider, for example, the rating of mental ability. All of the subjects in the study had had extensive experience in school which provided many opportunities for them to acquire a sound basis for judging their mental ability. School marks, teacher statements, feelings of success and failure, parent judgments, and peer evaluations are illustrative of the experiences which would help a person acquire an accurate self concept. Prior to counseling, 79 per cent of the subjects inaccurately rated their mental ability. After counseling, when they would have been given the knowledge necessary to rate themselves accurately, 68 per cent persisted in rating themselves inaccurately.

Comparison of Interest Profiles

The Kuder Preference Record—Vocational, Form C yields ten scores, each indicative of interests in a particular area. Counselors usually interpret individual scores in the light of the total profile.

Thus, one student's highest scores might be at the 85th centile in the clerical and musical areas, while another student's highest scores are in the same area but are equivalent to centiles of

60. In both cases, the counselor would probably interpret the highest scores as indicative of areas of greatest interest. But the interpretation in both would continue to the point of helping the client see the strength of his interests in relation to those of persons in the norm group. Essentially, profile interpretation is a process of discerning the strengths of interest in the ten areas measured by the Kuder in relation to each other and to the normative population. Because such interpretations were common practice among the counselors included in this study it appeared desirable to consider the accuracy with which students reported their obtained Kuder scores when taken together as a profile. To make the profile comparisons, the data were reprocessed using a technique of assessing similarity between profiles (5). The technique is illustrated by the following data concerning one student:

Column	Kuder Interest Areas									
	0	1	2	3	4	5	6	7	8	9
First rating	5	2	5	2	4	4	4	5	4	1
Test score	3	4	5	2	5	2	5	3	4	1
Difference	2	-2	0	0	-1	2	-1	2	0	0
Square of difference	4	4	0	0	1	4	1	4	0	0

In the "0" (Outdoor Interest) column, the "5" is the student's rating of his outdoor interest which he made before taking the test. The "3" is the quintile category based on his obtained test score. The "2" is the difference when the score is subtracted from the rating. And finally, the "4" is the square of the difference between score and rating. The data for the other areas were treated in a similar manner. A figure

of 18 is obtained by adding the numbers in bottom row above. The square root of 18, or 4.243, was used as this student's first-rating-profile score.

Another profile score was obtained for this student by comparing the interest ratings he made after counseling with his test scores. The technique used followed the pattern described above. This process yielded a second-rating-profile score. In the case of the student cited above, his second-rating-profile score was 3.464. Similarly, the data concerning all other students in the study were processed to yield for each student a first-rating-profile score and a second-rating-profile score.

Then the second-rating-profile score was subtracted from the first-rating-profile score. In the case of the student used as an example in the preceding paragraphs this computation yielded a positive number, .779 (i.e., $4.243 - 3.464 = .779$). The algebraic difference between the profile scores indicated whether the individual had moved toward greater congruency between the profile of ratings and the profile of scores, or toward less congruence. A positive difference between the profile scores indicated that the second profile of ratings corresponded more nearly to the obtained profile of scores than did the first profile of ratings. Conversely, a negative difference means that the first ratings were more consonant with scores than were second ratings. In fine, positive differences showed increased "accuracy" and negative differences a deterioration in "accuracy."

Separate frequency distributions of the differences between profile scores were made for the counseled, counseled-without-test, and not-counseled groups. The distributions ranged from high positive through zero difference to high negative.

The zero-difference category was not interpreted as indicating that the student did not change his profile of ratings. The zero difference merely indicated that the subject's second profile of ratings agreed neither less nor more closely with the profile of scores than did his first rating profile.

For the distributions of the differences between the first-rating-profile score and the second-rating-profile score the means and standard deviations were computed. Table 4 contains these data. The means

TABLE 4
MEAN AND STANDARD DEVIATION OF
DIFFERENCES BETWEEN PROFILE SCORES

Statistic	Coun- seled	Coun- seled With- out Test	Not Coun- seled
Mean of difference	.41	.26	.04
Standard error of mean	.09	.12	.07
Standard deviation	1.00	.95	.97
Number of cases	175	66	230

varied in magnitude in the expected manner; namely, the counseled group showed the most change in the direction of greater agreement, the not-counseled group the least change. A *t* test indicates that the means of the counseled and the counseled-without-test groups are significantly different from zero. In considering the mean difference of .41, one should not overlook the fact that the difference between score and rating could not exceed 4.00. Hence the difference is slightly more than 10 per cent of the largest possible difference. To test the significance of the difference between means, *t* ratios were computed. The *t* ratio between the mean of the counseled group and the mean of the noncounseled group was 3.52, significant at the .01 level. The *t* ratios between the mean of

TABLE 5

COMPARISON OF CERTAIN CHARACTERISTICS OF SUBGROUPS SELECTED ON BASIS OF ACCURACY OF RATING OF INTEREST

Variable	Counseled		Counseled Without Interest Test		Not Counseled	
	Increased	Deteriorated	Increased	Deteriorated	Increased	Deteriorated
Mean age	15.8	15.9	16.1	16.3	16.1	15.7
Mean school grade	10.4	10.6	10.7	10.6	10.3	10.3
Mean P.M.A. score	3.3	3.5	3.1	2.5	3.1	3.1
Mean number of tests interpreted	5.0	4.7	3.7	3.5	—	—
Number of: males	23	24	11	7	36	38
females	24	23	7	11	26	24
Number: desired counseling	45	45	15	12 ^a	25 ^a	30 ^a
did not desire counseling	2	2	3	3 ^a	28 ^a	23 ^a

^a Some students did not respond to the question, "Would you like an opportunity to talk with a counselor?" Therefore, these totals differ from others in the table. In instances where the number of nonrespondents was excessive, tests of significance were not applied.

the counseled-without-test group and the counseled and the not-counseled groups were found to be 1.03 and 1.63, respectively. Neither was statistically significant.

The data thus far reported have answered affirmatively the two questions spelled out in the section which dealt with the purpose of the study. They revealed that there was a significant difference between precounseling and post-counseling criterion measurements. The data also revealed that these observed differences were in the direction of more agreements after counseling. The stability of these differences was also demonstrated by the differences between the counseled, counseled-without-test, and not-counseled groups. The neat stair-step-
ping of mean differences in Table 4 from smallest for the non-counseled group to largest for the counseled group lends credence to the belief that the observed changes are in part at least a function of counseling.

Comparison of Factors

In this section data will be presented

concerning factors thought to be associated with changes measured by the criterion. They are: age, grade placement, intelligence, sex, number of tests interpreted, and desire for counseling.

In order to determine if the factors were associated with the criterion, the subjects were divided on the basis of the difference between the first-rating-profile score and the second-rating-profile score. The 27 per cent of the students with the largest positive differences (increase in "accuracy") and the 27 per cent of the students with the largest negative differences (deterioration in "accuracy") in the area of interests were identified in the counseled, counseled-without-test, and not-counseled groups. This process yielded the six subgroups.

For each subgroup the statistics summarized in Table 5 were obtained. When appropriate tests of significance were applied to differences between the "increased" and the "deteriorated" subgroups in the counseled, counseled-without-tests, and not-counseled groups, none of the differences was found to be statistically significant. Comparable results

TABLE 6
NUMBER OF PROBLEMS CHECKED ON THE *SRA Youth Inventory*

Inventory Areas	Counseled					Counseled Without Test					Not Counseled				
	Increased N=34		Deteriorated N=29		t ratio	Increased N=14		Deteriorated N=12		t ratio	Increased N=55		Deteriorated N=55		t ratio
	Mean	SD	Mean	SD		Mean	SD	Mean	SD		Mean	SD	Mean	SD	
My School	6.00	4.34	5.31	3.48	.68	6.64	3.87	4.92	3.17	1.20	7.25	4.40	6.44	4.50	.94
After High School	11.68	7.67	8.97	6.09	1.54	13.14	7.56	9.25	7.43	1.27	9.06	7.65	9.02	6.77	.68
About Myself	6.70	6.54	6.14	4.89	.38	8.86	5.29	6.75	8.70	.70	7.89	7.84	6.45	5.87	1.07
Getting Along with Others	6.79	6.42	6.93	4.50	.10	8.29	6.17	4.67	5.02	1.59	6.96	6.36	6.22	6.22	.61
My Home and Family	4.18	6.12	3.52	5.35	.45	4.36	3.29	2.50	3.23	1.05	5.11	6.46	3.06	5.85	.97
Boy Meets Girl	3.91	3.93	3.93	4.30	.02	3.57	3.81	2.92	1.80	.54	3.56	4.77	3.24	3.55	.40
Health	2.62	2.29	2.79	2.59	.27	2.57	2.82	2.33	1.97	.25	3.05	3.14	2.65	2.32	.74
Things in General	4.18	4.34	3.66	3.17	.54	4.21	4.99	4.00	3.51	.38	3.93	4.37	3.55	4.74	.44
Basic Difficulty Score	12.41	11.24	11.66	8.58	.30	14.14	7.94	12.50	13.50	.35	15.05	12.76	12.65	10.20	1.08

were found when mental ability and mechanical aptitude were similarly treated.

On the basis of these data, it was concluded that the improvement or deterioration in the "accuracy" of students' self-ratings apparently was not a function of their age, school grade, mental ability, sex, desire for counseling, or number of tests interpreted during counseling. It appeared, therefore, that the criterion employed in this study will be useful in the further study of counseling because it was not influenced by gross factors over which counselors have little control. This conclusion led to a desire to learn even more about factors associated with the criterion. One attempt to study such factors will be reported in the next section.

Comparison of Problems

The *SRA Youth Inventory* appeared to be an instrument which might be used to identify some of the more subtle characteristics which distinguished those students who show more agreements between scores and ratings after counseling than before. The *Inventory* is a list of 298 problems of which the testee checks those which apply to him. The more problems he checks the higher his score.

Did students who showed more agreements have more or fewer problems than those who showed fewer agreements after counseling? The answer to this question was sought in a comparison of the *SRA Youth Inventory* scores of 100 students who showed the greatest increase in agreements with 99 students who showed the greatest deterioration in the agreement between their profiles of interest scores and ratings. These students were drawn from the upper and lower 27 per cents of the distribution of differences between first- and second-rating-profile scores previously described. The mean inventory score of the "increased" group was 46.92 with a standard deviation of 30.32. The "deteriorated" group had a mean inventory score of 41.72; the standard deviation was 29.64. The *t* ratio for the difference between these means was 1.22, not significant. Hence, the question asked above was answered: there was no difference between the groups in terms of the total number of problems checked on the *Inventory*.

The next step was to determine whether or not there was a difference between the "increased" and "deteriorated" groups in the number of problems checked in *specific areas*. The *SRA Youth Inventory*, as previously pointed out, yielded scores in eight areas of adjustment. In order to further refine the analysis, the groups were broken down into counseled, counseled-without-test, and not-counseled subgroups. The mean number of problems checked by each subgroup for each area is presented in Table 6. In Table 6 the *t* ratios based on the differences between means are also given. None was found to be statistically significant. Hence, it was concluded that the *SRA Youth Inventory* scores did not identify those students who showed more agreements between their profile of scores and ratings after counseling than before.

The two comparisons based on number of problems checked did not take into account the

possibility that certain particular problems might be checked by the "increased" group and others by the "deteriorated" group. The previously discussed analyses were based simply upon number of problems checked. Hence an item analysis was undertaken. The groups were divided into random halves. The resulting groups were (a) Increased I, $N = 50$; (b) Increased II, $N = 50$; (c) Deteriorated I, $N = 49$; and (d) Deteriorated II, $N = 50$. The item responses of the Increased I and Deteriorated I groups were tabulated and percentage of response for each group for each item was computed. If the difference between the percentages for the two groups was significant at the .05 level, the item was retained.

There were 35 of the 298 items in the test retained. Of these, 32 were keyed for increase and 3 keyed for deterioration. These items were cast into a scoring stencil and all answer sheets for the test were rescored. A single score was obtained by subtracting the deterioration-keyed items from the increase-keyed items. The mean scores and standard deviations were computed. These data are presented in Table 7. The differences between the means of the Increased I and Deteriorated I groups yield a t ratio of 7.58, highly significant. This was expected since the scoring key was based on these groups. In an attempt to cross validate, the 35-item scoring key was applied to the answer sheets of the subjects included in the Increased II and Deteriorated II groups. The data in Table 7 con-

tion of the number or kind of problems the student checked.

SUMMARY AND DISCUSSION

The over-all purpose of this investigation was to study a proposed criterion for the evaluation of counseling, namely, *change in agreement between self-rating and score*. It is appropriate to ask, "What was learned about the criterion?" Details in the answer to this question can be found in the preceding pages. They may be summarized briefly. The criterion data varied in the expected direction, that is, the findings followed the logic of improved self-knowledge after counseling. From this point of view the criterion which was adopted appears to hold promise as a useful indicator of counseling effectiveness.

Its potential usefulness was further substantiated by the findings that the criterion appeared not to be affected by age, sex, school grade placement, intelligence, or desire for counseling. These are factors over which a counselor has no control (except by limiting his practice to specified types of clients). The important implication of these findings is that the criterion variable is independent of such factors which are extraneous to the counseling process *per se*.

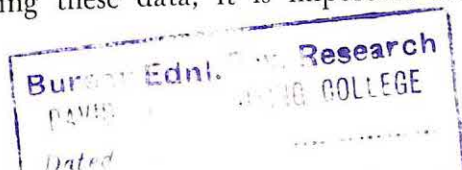
A related finding was that neither the number nor the kind of "problem" which counselees verbalized on the problem checklist was associated with the criterion variable. If this result were accepted at face value, it would do violence to beliefs which counselors have about their relative effectiveness with clients having "minor" as opposed to "major" problems of adjustment, or with clients having certain kinds of problems (e.g., vocational choice as opposed to feelings of inferiority). In considering these data, it is important to

TABLE 7
MEAN AND STANDARD DEVIATIONS ON SCORES
DERIVED FROM ITEM-ANALYSIS KEY
(Cross-Validation Sample)

Group	N	Mean	Standard Deviations
Increased I	50	8.00	4.56
Deteriorated I	50	2.51	2.20
Increased II	50	4.76	2.84
Deteriorated II	49	5.18	3.54

cerning these two groups indicated that the 35-item key did not hold up. When a t test was applied to the difference between the means of these two groups, it yielded a t ratio of .65, not significant.

These analyses of the *SRA Youth Inventory* data suggested that the increased-deteriorated classification was not a func-



keep in mind that the criterion involved ratings in the areas of interests and ability. It was not concerned with so-called problems of personal adjustment, the forte of the problem checklist used in this investigation. But even in areas of the problem checklist closely related to interests and ability, no statistically significant association with the criterion was noted.

This brief synopsis of the findings of this investigation can serve as a starting point for considering next steps.

The criterion explored in this investigation is an "immediate" one, i.e., it is available within a short time after the experiment is completed. But its relationship to the long-range objectives of counseling has not been established. Herein lies a great weakness. If one were to argue that self-knowledge is the object of counseling, the criterion studied might be presumed to be a reasonable measure. However, few counselors, if any, would be satisfied with self-knowledge as an end in itself. The common expectation is that

self-knowledge should influence counselee behavior. Here the criterion is silent, for it does not measure behavior. And in this silence is found an important gap in the usefulness of the criterion, as well as a suggestion for additional research. In future research it might be possible, for example, to discover that the criterion is associated with more intelligent, more constructive, or more adaptive postcounseling behavior. If such a discovery were made, then the criterion would acquire sociological significance.

Another fact must be considered. Some students who were not counseled showed as much change on the criterion in the direction of more "accurate" ratings as did students who were counseled. Conversely, some counsees deteriorated in the "accuracy" of their ratings as much as some not-counseled students. What factors are associated with change as measured by the criterion? The design of the present experiment was such that only counseling as total process was studied. Are there variations within the process from counselor to counselor, counselee to counselee, or problem to problem which account for the observed differences in the criterion variable? If so, are these factors also at work in different form among not-counseled students? These are crucial questions for which no ready answers are available.

REFERENCES

1. BENNETT, G. K., & FRY, DINAH E. *Test of mechanical comprehension, Form BB*. New York: Psychological Corp., 1941.
2. BUROS, O. K. (Ed.) *The nineteen forty mental measurements yearbook*. Highland Park, N.J.: The Mental Measurements Yearbook, 1941.
3. BUROS, O. K. (Ed.) *The third mental measurements yearbook*. New Brunswick, N.J.: Rutgers Univer. Press, 1949.
4. BUROS, O. K. (Ed.) *The fourth mental measurements yearbook*. Highland Park, N.J.: Gryphon Press, 1953.
5. CRONBACH, L. J., & GLESER, GOLDINE C. Assessing similarity between profiles. *Psychol. Bull.*, 1953, 50, 456-473.
6. KUDER, G. F. *Kuder preference record-vocal*
7. REMMERS, H. H., & SHIMBERG, B. *SRA youth inventory, Form A*. Chicago: Science Research Associates, 1951.
8. ROGERS, C. R., & DYMOND, ROSALIND F. (Eds.) *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954.
9. THURSTONE, L. L., & THURSTONE, THELMA G. *SRA primary mental abilities: intermediate*. Chicago: Science Research Associates, 1949.
10. WERT, J. E., NEIDT, C. O., & AHMANN, J. S. *Statistical methods in educational and psychological research*. New York: Appleton-Century-Crofts, 1954.

(Accepted for publication March 11, 1957)

Aggression in Fantasy and Overt Behavior¹ARTHUR R. JENSEN²*Teachers College, Columbia University*

THIS PAPER is concerned with the relationship of Thematic Apperception Test responses to overt behavior. A review of the research on this subject (9, ch. 1) suggests the following generalizations:

a. The thematic content per se of TAT stories seems to be related very slightly, if at all, to general behavioral traits or dispositions. But thematic content does seem to be related to temporary, situationally induced behavior, affective states, and drives.

On the basis of correlations between the fantasy themes of 40 adolescent boys and girls, and teachers' ratings of behavioral characteristics, Symonds concluded that the relationships between these two sets of variables were "insignificant and negligible" (19, p. 322). Studies by Pittluck (15) and Gluck (7) found no relationship between fantasy aggression and general overt aggressiveness. On the other hand, Bellak (3) showed that the number of aggressive words in TAT stories increased when the storyteller was insultingly criticized for the low quality of his stories. Essentially the same finding was made by Feshbach (6), who also showed that Ss who had a fantasy outlet for their induced aggression (by writing TAT stories) expressed less aggression in a sentence completion test designed to get at the Ss' hostility toward the insulting examiner. Clark (1) showed that male Ss who presumably had been sexually aroused by being shown photographs of nude women had fewer themes of sex in their TAT stories than did control Ss, while another group of Ss under the disinhibiting effects of alcohol produced more themes of sex in their TAT stories after having been shown pictures of nude women. Lindzey (11) demonstrated that extrapunitive behavior on the part of the hero in TAT protocols increased significantly following failure in a social situation. McClelland et al. (12) have reported substantial positive correlations between *n* Achievement in TAT fantasies and several measures of behavioral achievement when the Ss were tested in an achievement-oriented situation. Sanford (16, 17), and Atkinson and McClelland (1) have shown that Ss produce more food-related responses to the TAT when they are made hungry by food deprivation.

b. The expressive or behavior-sample aspects of the TAT responses seem to be correlated with overt behavioral traits and would therefore seem to be a more valid basis for predicting overt behavioral traits than are the thematic or fantasy aspects of the TAT.

An example of this is Balken and Masserman's study (2), showing language and stylistic differences between obsessives, hysterics, and anxiety states. On the basis of a study correlating various aspects of TAT responses with various personality variables, Hartman stated: "Attention in the past has been centered upon thematic analysis in TAT interpretation. Formal characteristics of TAT responses should be given increased emphasis inasmuch as they can be more objectively determined and may, particularly in application to group testing, be more revealing of certain aspects of personality" (8, p. 35).

Problem

The present study examines both the

¹ This article is based on a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, under the Joint Committee on Graduate Instruction, the Columbia University. The author is indebted to Dr. Percival M. Symonds, under whose direction and encouragement this study was undertaken.

An abstract of this article was read at the annual meeting of the Eastern Psychological Association, Atlantic City, New Jersey, March, 1956.

² Now at the Institute of Psychiatry (Maudsley Hospital), University of London.

thematic and the behavior-sample aspects of the TAT and their relation to overt aggressive behavior as a general disposition or trait. Aggression was the variable chosen for study because it lends itself particularly well to this kind of investigation; it is more easily identified and can be more reliably rated in both the TAT stories and in the S's behavior than can many of the other personality variables purportedly revealed by the TAT. A review of the existing research led to the formulation of the following questions, in terms of which the present study was planned:

1. Is there a *direct* relationship between aggression in TAT fantasy and in overt behavior? In other words, do Ss who show much overt aggression have more themes of aggression in their TAT stories than do Ss who show little aggressiveness?
2. Is there an *inverse* relationship between fantasy aggression and overt aggression? Both Symonds (19) and Sanford (18) have suggested that needs which are repressed in overt behavior are more likely to find expression in fantasy, and that needs which are worked out in behavior are less likely to be expressed in fantasy.
3. Does the degree to which *punishment* is associated with aggression in fantasy permit prediction of overt aggressiveness? Some clinicians believe that when in fantasy punishment is associated with aggression, the S is less apt to be overtly aggressive. A study by Mussen and Naylor (14) lends support to this hypothesis.
4. The same idea has been put forth with respect to the amount of *defense* against aggression anxiety in the TAT. Is an S who denies the aggressive implications of a TAT picture or who has to rationalize or tone down the aggressive implications of his own story less apt to be overtly aggressive than an S who does not defensively modify his aggressive fantasies? Pittluck's research (15) suggests an affirmative answer to this question.
5. Do TAT stories when viewed as a *behavior sample* show a relationship to the S's overt behavior in a wide variety of other situations? For example, would aggressively delinquent boys reveal in their TAT productions the socially defiant, rebellious attitude which characterizes their behavior in other situations? Delinquent or antisocially aggressive behavior may be interpreted partly as a rejection and defiance of certain social standards (5). The TAT would thus

be expected to elicit from delinquent boys behavior which violates the taboos of the examiner or of the setting in which the TAT stories are produced. In other words, the TAT production may be viewed as a direct sample of the S's behavior and not simply as a fantasy which indirectly reflects the S's needs and conflicts. According to this hypothesis boys who exhibit socially unacceptable forms of aggressive behavior in school would give a sample of their "unacceptable" behavior in the TAT by using tabooed language, themes of sex, tabooed forms of sex, and unusually repulsive or tabooed forms of violence, such as mutilation of the victim. These forms of expression may be regarded as aggressive acting-out in the TAT situation and are therefore direct behavior samples.

On the basis of these questions there were formulated the specific hypotheses which could be tested by comparing the TAT productions of three groups of Ss differing widely with respect to overt aggressive behavior. Each group was relatively homogeneous as regards aggressive behavior, and the Ss in each group were selected as relatively extreme representatives of a particular type of behavior. It was believed that these conditions would allow a maximal opportunity for the relevant TAT variables to manifest significant differences. It was considered important that all the Ss be of approximately the same age, be selected from the same setting, and be tested under the same conditions. These requirements were fulfilled. The three groups, composed of high school boys, were labeled *Aggressive-Bad* (socially unacceptable overt aggression); *Aggressive-Good* (socially acceptable overt aggressiveness); and *Passive* (extremely lacking in all overt aggressive qualities). The method of selecting the Ss is described in the following section.

Hypotheses

1. The Aggressive-Bad group has the most TAT aggression; the Passive group has the least. (Direct relationship.)
2. The Aggressive-Bad group has the least

TAT aggression; the Passive group has the most. (Inverse relationship.)

3. The proportion of mild forms of TAT aggression (Scores 2-6) is smallest for the Aggressive-Bad group. (Qualitative relationship.)

4. The proportion of stories in which the hero is the victim of aggression to those in which the hero is the aggressor is smallest for the Aggressive-Bad group, largest for the Passive group.

5. The Aggressive-Bad Ss have fewer themes of suicide than either of the other groups.

This is based on the belief that suicide represents aggression turned against the self and that nonaggressive Ss tend to be more intropunitive and develop reaction formations against the overt expression of aggressive impulses.

6. The Aggressive-Bad group has the smallest proportion of punishment relative to aggression in their TAT stories; the Passive group has the largest proportion. (Dynamic relationship.)

7. The Aggressive-Bad group has the smallest proportion of defense relative to aggression; the Passive group has the largest proportion. (Dynamic relationship.)

"Defense" refers to an evasion used by the storyteller, presumably to avoid aggression anxiety. The aggressive fantasy is not expressed as "raw" aggression, but is rationalized, made socially acceptable, or accidental, or is qualified in such a way that its aggressive implications are lessened or hidden.

8. The Aggressive-Bad group has the largest proportion of aggressive stories containing "raw" aggression, i.e., aggression free from association with punishment (or guilt) or defense. This is a corollary of Hypotheses 6 and 7. (Dynamic relationship.)

9. The Aggressive-Good and Passive Ss have more themes of natural death than the Aggressive-Bad Ss. (Dynamic relationship.)

This is derived from the belief that natural death occurring in a TAT story represents a disguised expression of aggression and is thus psychologically similar to the above-mentioned defenses.

10. The Aggressive-Bad group has more themes of sex, tabooed sexual content, tabooed language, and tabooed or "shocking" forms of violence than either of the other groups. (Behavior-sample relationship.)

This is based on the assumption that the expression of sexual and tabooed themes in the TAT stories, especially when they are written in a school setting, represents a lack of inhibition about expressing socially disapproved thoughts and actions. Tabooed language and tabooed sex are also forms of verbal aggression—perhaps the only forms of verbal aggressive behavior that the test situation permitted. (A more open form of aggressive behavior would have resulted in

the Ss' dismissal from the testing room.) When these tabooed ideas can "break through" into written expression, it would seem to indicate both a certain preoccupation with the socially tabooed and a flaunting defiance of customary decency and propriety. The manifestation of these attitudes in the general school behavior of the Aggressive-Bad Ss was probably among the reasons many of them were selected by their teachers for this experiment.

PROCEDURE

Subjects

Three types of Ss were desired, each showing different behavioral characteristics with respect to overt aggressiveness. As a convenience, the three types were labeled *Aggressive-Bad*, *Aggressive-Good*, and *Passive*.

The *Aggressive-Bad* Ss were those whose aggressiveness got them into trouble at school. They were overtly aggressive in disruptive, socially unacceptable ways.

The *Aggressive-Good* Ss were those whose aggressiveness found socially acceptable expression in athletics, leadership, competition, and other nondisruptive activities.

The *Passive* Ss were those so lacking in aggressive qualities as to be shy, quiet, unobtrusive, unassertive, and diffident.

Ss of these three types were selected in an all-boy high school with over a thousand students, most of whom live in an industrial community. Every teacher in the school was provided with instructions for selecting these Ss. Each was asked to submit the names of five boys he knew from personal experience during the current school year (this was done in May) who most fully met the description of the specified types. Every teacher was asked to rank these Ss for degree of conformity to the behavioral traits described and also to fill out check lists of various descriptive criteria with respect to kinds of aggressive behavior, indicating those characteristics which best described his reasons for selecting each S.

Fifty-one teachers submitted a total of 756 names. This represents only 433 students, since some were submitted by as many as six teachers. There were disagreements in classifying 25 of the students. These were immediately eliminated as possible Ss for the present study.

From the remaining 408 students a smaller number was selected for testing. The Ss were selected first on the basis of the number of teachers submitting their names. The second basis for selection was the rank assigned to the

S by the teacher. The behavior check lists were used to ensure a high degree of homogeneity within each of the groups with respect to types of aggressive behavior. Items frequently checked by the teachers for Ss in each group were:

Aggressive-Bad. Disruptive, unruly, bad temper, defiant of authority, headstrong, boisterous, "tough," surly, destructive of property, aggressive bravado.

Aggressive-Good. Active, energetic, helpful, leader, likes competition, "carries the ball," self-assertive, acceptable verbal aggression (e.g., debate), seeks recognition, engages in competitive sports.

Passive. Quiet, retiring, meek, unassertive, withdrawing, diffident, timid, does not actively participate, submissive, a follower, "shrinking violet."

It is believed that relatively homogeneous groups of the most extreme and representative Ss of each type were obtained by this procedure. All the Ss were between 15 and 17 years of age. Fortuitous circumstances at the time of testing prevented the Ns in each group from being equal (Aggressive-Bad = 25, Aggressive-Good = 22, Passive = 27.)

Selection of TAT Pictures

A pilot study was conducted in a comparable all-boy high school in another city in order to determine the most effective and efficient method of administering the TAT and the most suitable pictures for the purpose of this study. The Ss in the pilot study were students in four different English classes (total $N = 90$). Four different methods of group administration of the TAT were tried. Analysis of these data led to the following conclusions: (a) the Ss were most productive when required to write stories to *five* of the TAT pictures during the class period of 50 minutes; (b) there were no systematic differences in the stories (as scored for this study) between 10th-, 11th-, and 12th-grade Ss; (c) all the stories were scored on the main variables of this study and from these data the TAT pictures were ranked according to their power to elicit these themes. The

theme of aggression was weighted most heavily in determining the rank order of the pictures. The ten most fruitful cards, in the order of their effectiveness, are listed below. These were used in the main study.

- 18BM ... Man—hands clutching him from behind.
- 13MF ... Woman lying on bed—man standing near.
- 17GF ... Woman standing on bridge over water.
- 18GF ... A woman has her hands squeezed around the throat of another woman.
- 3GF ... Woman standing in doorway with downcast head.
- 4 ... Young woman clutching shoulders of young man who is turning away.
- 3BM ... Boy huddled on floor against couch.
- 9GF ... From behind a tree a young woman watches another woman running along a beach.
- 20 ... Man standing under street light.
- 15 ... Gaunt man standing among grave-stones.

Administration of the TAT

The Ss were tested in groups of about 25 Ss of all types (Aggressive-Bad, Aggressive-Good, and Passive). Thus there were three test sessions. Ss sat in alternate seats and were well proctored. Each session lasted 90 minutes and all Ss wrote ten TAT stories within that time. A shortened and simplified version of the standard Murray instructions (13) was printed on the cover of the test booklets in which the Ss wrote their stories; these instructions were also read aloud by the examiner at the beginning of the session.

Three sets of TAT cards were used; the examiner exchanged each S's picture as each story was completed, or the Ss themselves exchanged pictures. Thus each S could write at his own speed, though to ensure that everyone would finish in one session, the examiner announced at 8-minute intervals that a certain number of stories should have been completed. An examination of the mean scores on the TAT variables for

each of the three test sessions (9, p. 96) showed that if there was any "leakage" of information about the TAT from Ss in the first session to the Ss in the second and third sessions it had no effect on the results of the experiment. There were no consistent or significant differences between the three sessions.

Scoring the TAT

A special scoring system was devised for the purpose of this study. The key variables are *Aggression*, *Punishment* for the aggression, and *Defense* against aggression-anxiety. A distinction was made between whether the hero (central character) of the story is the aggressor or the victim of aggression. Since a more refined means of classifying aggression and punishment was desired, these variables were subdivided and assigned "scores" according to their intensity.

All the other variables are indirectly related to aggression. They were scored simply on the basis of their presence in the story. *Suicide* is aggression against the self. The occurrence of *Natural Death* in a story may be a disguised expression of aggression. *Sex* and *Tabooed Language* may have aggressive connotations or may be an expression of aggression in the test situation. Also they may indicate weak inhibition of socially disapproved forms of expression. *Tabooed Violence* is a particularly primitive and sadistic expression of aggression. The scoring instructions, given below, were made as complete, explicit, and unequivocal as possible, so that by following them carefully even persons not trained in the TAT can achieve close agreement in scoring.

Instructions for Scoring the TAT

Score only what is explicitly stated in the story. Do not make any inferences about what might be implied.

Aggression Score

Hero Aggressor. The main character of the story is responsible for the aggression; he (or she) is the one who can be called the aggressor. For determining the main character, see Murray's criteria for distinguishing the hero (13, p. 6).

Hero Victim. When the main character of the story is aggressed against by another character in the story, the aggression is scored in the *Hero Victim* category, since the hero is not the one who commits the aggressive act but is the victim of the aggressive act.

Indeterminate. When there is aggression in the story and it is difficult or impossible to decide if the story has a main character, the aggression should be scored in the *Indeterminate* category.

- 0 Inert, no action, mere contemplation; neutral or nonaggressive action; no indication of aggression in any form.
- 2 A character in the story dominates or controls another, takes self-initiated action, seems to do things with a purpose of his own.
- 4 Character in the story is actively working, struggling, striving, willfully overcoming obstacles, resisting pleas of others, daring exploits, being actively brave, etc.
- 6 Verbal aggression; swearing at another, "telling off" another, threatening another; verbal coercion.
- 8 Physical aggression not resulting in death and which usually is not expected to cause death: fighting, struggling with another, hitting, causing injury, physically subduing, rape, robbery, and deliberate destruction of property.
- 9 Physical aggression not resulting in death but which usually results in death, such as shooting another person, choking someone, knifing, inflicting any kind of injury which seems severe enough to result in death. (This score differs from 10 in that killing or death, or the intent to kill or cause death, are not explicit.)
- 10 Physical aggression resulting in death: murder, killing. (Killing or intent to kill must be explicit.)

Punishment Score

There can be a Punishment score only if the Aggression score is 6 or greater. The punishment is always for aggression in the story. If a character receives what seems to be punishment and there is no indication of previous aggression on his part, then what seems like punishment should be scored under *Hero Victim*. For example, the story: "They caught this man and hanged him," may seem like punishment, since being hanged is usually a form of punishment,

but since we do not know explicitly from the story that this man is being punished for an aggressive act, and since it is explicit that he is being aggressed against, we should put a score of 10 in the *Hero Victim* category.

- 0 No punishment of any kind.
- 2 Feelings of guilt, sorrow, depression, regret, or bad conscience. (Score only if explicit, as, e.g., "He feels sorry he did it," "He feels bad," etc.)
- 4 Nonphysical punishment from the environment: deprivation, rejection by others, loss of loved persons or objects.
- 6 Physical punishment: bodily harm, misfortune, caught by police, imprisonment.
- 8 Physical punishment resulting in death: execution. Suicide when it is self-punishment for an act of aggression. (Also scored under Suicide.)

Defenses

Defense is scored only when the Aggression score is 6 or greater. It represents a defense against aggression-anxiety. Defense is never scored unless one of the following is true in connection with the scored aggression. (The Defense score applies to Suicide as well as to Aggression.)

- a. Storyteller rejects an act of a figure; voices disapproval of it.
- b. Storyteller denies something. ("That couldn't be a gun." "She doesn't want to hurt her.")
- c. The aggressive activity is excused or rationalized. Aggression is put in a socially acceptable form (accident, self-defense, sports activity, warranted punishment, fighting or killing in war, police defending others, killing in the line of duty, etc.).
- d. The aggressive activity is a dream, thought, wish, fantasy, or plan of the story character, but it is not completed through action.
- e. The action is described as past or future. (This refers to the distant past or future; that is, the action does not take place within the time-span of the particular story.)
- f. The activity is qualified; maybe, perhaps, might be, probably, as if, doubt about the nature of the activity, or a nonaggressive alternative to the activity is suggested. ("This woman is either choking the other one or she's helping her because she's sick.")
- g. Noncompletion of aggression planned by the fantasy character. ("He goes out to kill this man, but then he changes his mind.")
- h. Displacement of aggression on to nonhuman objects. ("Man beats his dog." "Man kicks a chair, smashes a vase.")
- i. Cause of death is unknown, but it is not clearly due to natural causes. When there is

death in a story and it is not explicitly the result of aggression or of natural or accidental causes—in short, when there is no evidence in the story for inferring the cause of death—score Aggression 10 and Defense 1. ("The man enters the room and finds a dead body on the floor.")

Other Variables

The following are scored simply on the basis of their presence in a story.

Natural Death. A character dies, or has just died, from presumably natural causes.

Suicide. A character takes his own life, or is in the act of doing so.

Tabooed Violence. Scored in addition to an Aggression score of 8, 9, or 10 when the aggression is especially brutal, bloody, gory, repulsive, primitive, or sadistic. Also, mutilation of the victim.

Sex. Heterosexual petting or intercourse. Petting is scored as Sex only if it involves physical contact with the breasts or genitalia or explicitly arouses sexual impulses.

Score Sex only if it is actually part of the story and not merely suggestive. The language used is not important.

Tabooed Sex. Tabooed Sex refers to any explicit sexual activity which does not come under the heading of Sex, i.e., any sexual activity other than heterosexual petting or intercourse. Examples that occur in these stories: homosexuality, incest, masturbation, bestiality, necrophilia, fellatio, cunnilingus, and other perversions. In the stories these technical terms are not used. It is assumed that the scorer knows the slang terms.

Tabooed Language. Most of the tabooed language is in reference to sex. Do not score mere swearing as Tabooed Language. Only words that are generally considered "dirty" are classed as taboo. Usually these are various synonyms for sexual intercourse, masturbation, sexual anatomy, and the like.

Relationships between the various TAT variables were determined by dichotomizing all distributions of scores for each variable as close to the median as possible and computing the contingency coefficient as the measure of relationship (Table 1).

Reliability of Scoring

The Ss' names were removed from the stories, which were given code numbers,

TABLE 1
CONTINGENCY COEFFICIENTS BETWEEN VARIABLES

Variable	1	2	3	4	5	6	7	8	9	10
1. Aggression.....44†	.34*	.25	-.12	.14	.33*	.22	.14	.16
2. Punishment.....05	.17	.17	.32*	.19	.17	-.08	.38†
3. Defense.....	-.27	.22	.21	-.11	-.12	.00	.17
4. Sex.....	-.35*	.38†	.37†	.46†	.66†	.19
5. Natural Death.....05	-.16	-.12	-.24	.03
6. Suicide.....15	.00	.06	.04
7. Tabooed Violence....38†	.37†	.16
8. Tabooed Sex.....51†	-.19
9. Tabooed Language....09
10. Number of Words....

* Significant at or below the .05 level.

† Significant at or below the .01 level.

so that the experimenter did not know while scoring the stories to which group the Ss belonged. Another person, working independently without any knowledge of the purpose of the study or the nature of the Ss, scored all the odd-numbered protocols. The percentage of agreement in scoring the various TAT categories ranged from 86% to 99%; agreement for all categories combined was 95%.

RESULTS

Aggressive Fantasy Content

The distribution of Total Aggression Scores for the combined groups was dichotomized at the median and the significance of differences between the three groups of Ss was tested by means of the chi-square statistic. Only when the differences between all three groups reached the .05 level of significance were differences between pairs of groups tested for significance. (To facilitate direct comparisons of the frequencies in the three groups, all Ns in Tables 2, 3, and 4 have

been made directly comparable. In order not to create the distorted impression of large differences that would be given if percentages were used, these have been multiplied by .25, so that all frequencies in Tables 2, 3, and 4 may be viewed as if there were 25 Ss in each group. The chi squares were, of course, computed from the original frequencies.)

From Table 2-A it may be seen that both *Hypotheses 1* and *2* were not substantiated. The Aggressive-Bad and Passive groups did not differ significantly in total TAT Aggression. The Aggressive-Good group had significantly (.05 level) less TAT aggression than the Aggressive-Bad group, but did not differ significantly from the Passive group.

The relationships between the groups are essentially the same when comparisons are made of Aggression scores derived only from stories in which the Hero is the aggressor (Table 2-B).

Hypothesis 3 was not substantiated. The groups did not differ significantly in the proportion of stories containing Mild Aggression (scores 2-6) to stories containing Strong Aggression (score 8-10). Dichotomizing at the proportion of .30 maximizes the differences between the groups and yields the largest chi square of all possible dichotomies. Yet the differences between the groups did not attain significance (Table 2-C).

TABLE 2
AGGRESSIVE FANTASY CONTENT

TAT Variable	Cutting Point	Number of Subjects ^b			Chi Square	Chi Square ^a		
		Agg.-Bad	Agg.-Good	Passive		Agg.-Bad vs. Agg.-Good	Agg.-Bad vs. Passive	Agg.-Good vs. Passive
(A) Total Aggression Score	Above Median	16	7	14	6.82 ^f	4.42 ^f	.31	1.65
	Below Median	9	18	11				
(B) Hero Aggressor Score	Above Median	17	9	11	5.18			
	Below Median	8	16	14				
(C) Proportion of	Mild Agg. ^c							
	Strong Agg. ^d							
(D) Proportion of	Hero Victim							
	Hero Aggressor							
(E) Suicide	1 or more stories	15	16	8	5.71			
	none	10	9	17				

^a With Yates' correction.

^b Though the Ns of the three groups were actually 25, 22, and 27 respectively, they have here been "equalized" to an N of 25 in each group to facilitate direct comparisons. The chi squares are based on the actual Ns.

^c Aggression scores 2-6.

^d Aggression scores 8-10.

^e This cut gives the largest chi square of all possible cuts.

^f $p < .05$.

Hypothesis 4 was not substantiated. The groups did not differ significantly in the proportion of stories in which the Hero was the victim of aggression to stories in which the Hero was the aggressor. Dichotomizing at the proportion of .50 yields the largest chi square of all possible dichotomies, yet it falls far short of significance (Table 2-D).

Hypothesis 5 was not substantiated. The groups did not differ significantly in the number of stories containing the theme of Suicide. Dichotomizing simply on the basis of presence or absence of Suicide in the entire set of 10 stories comes closest to dividing the distribution of Suicide scores at the median; no other dichotomy yields a larger chi square for these data (Table 2-E).

Hypothesis 6 was not substantiated. The three groups did not differ significantly in the proportion of Punishment relative to Aggression in their TAT stories (Table 3-A). When the groups are compared for themes of Punishment, not on a proportional basis, but on the basis of absolute number of Punishment themes in their stories, the differences between the groups are even smaller.

Hypothesis 7 was not substantiated. The three groups did not differ significantly in the proportion of Defense relative to Aggression in their TAT stories (Table 3-B).

Hypothesis 8. Since the findings with regard to Hypotheses 6 and 7, while not statistically significant, were in the predicted direction, and

since both Punishment and Defense are psychologically similar in that they may be interpreted as indicative of the S's disapproval of his aggressive fantasy, Punishment and Defense were regarded as equivalent in Hypothesis 8. It may be argued that if an S uses one means of "disapproving" his aggressive fantasy, he need not use another means in the same story. The groups were therefore compared on the basis of the proportion of *either* Punishment or Defense relative to aggression. Themes of aggression in which there is neither Punishment nor Defense may be called "raw" or unmodified aggressive fantasy. When the distribution of raw Aggression scores is dichotomized as near the median as possible, it is seen that the groups differ significantly (Table 3-C). The Aggressive-Bad group had more raw aggression in their stories than either the Aggressive-Good or Passive groups, which did not differ significantly from each other. Thus, *Hypothesis 8* was substantiated.

Hypothesis 9 was only partially substantiated. The Aggressive-Good group had significantly more themes of Natural Death than either the Aggressive-Bad or Passive groups, which did not differ significantly from each other (Table 3-D).

Behavior-Sample Aspects

Hypothesis 10 was amply substantiated. All the behavior-sample variables discriminated between the groups at the .01 to .001 level of significance (Table 4).

TABLE 3
MODIFIERS OF AGGRESSIVE FANTASY

TAT Variable	Modifier	Number of Subjects ^b			Chi Square	Chi Square ^a		
		Agg.-Bad	Agg.-Good	Passive		Agg.-Bad vs. Agg.-Good	Agg.-Bad vs. Passive	Agg.-Good vs. Passive
(A) Aggression ^c	With Punishment ^e	9	12	7	2.05			
	No Punishment	15	11	16				
(B) Aggression ^c	With Defense ^e	13	16	18	2.64			
	No Defense	11	8	6				
(C) Aggression ^d	With Punishment or Defense ^f	4	15	12	10.14 ^h	7.43 ^h	4.54 ^h	.22
	No Punishment or Defense	16	7	9				
(D) Natural Death	Present ^e	6	17	8	10.44 ^h	7.54 ^h	.72	4.58 ^g
	None	19	8	17				

^a With Yates' correction.

^b See footnote b in Table 2.

^c Based only on Ss having one or more stories containing Aggression, thus comparing the groups on the basis of the proportion of Punishment relative to Aggression.

^d Based only on Ss having two or more stories containing Aggression, thus comparing the groups on the basis of the proportion of Punishment or Defense relative to Aggression.

^e In one or more stories.

^f In two or more stories.

^g $p < .05$.

^h $p < .01$.

Because of the markedly skewed distribution of scores on each of these variables, the group comparisons were made on the basis of the number of Ss in each group having a given number of stories containing the particular TAT variable. In some cases a true picture of the differences between the groups is obtained only when the distribution of scores is dichotomized at more than one point. This was the case with Sex and Tabooed Sex.

On every variable in Table 4 the Aggressive-Bad group was significantly ahead of the Aggressive-Good and Passive groups, which did not differ significantly from each other. The one partial exception to this, shown in Table 4-A, is Sex in one or more stories. Here the Aggressive-Bad and Aggressive-Good groups did not differ significantly from each other, but both differed significantly (.01 to .001 level) from the Passive group, which had very few themes of Sex. However, when the groups are compared on the basis of the number of Ss having two or more stories containing Sex themes, the Aggressive-Bad group is far in the lead, while the Aggressive-Good and Passive groups do not differ significantly. The reason for this becomes obvious when the groups are compared on the individual TAT cards. The Passive Ss, apparently afraid of doing anything that they think might incur

disapproval, did not respond with sexual themes even to Card 13MF, which usually elicits sexual themes. Indeed, sexual themes may be regarded as appropriate for this picture of a man standing beside a half-nude woman lying on a bed. The majority of sexual themes in the Aggressive-Good group were given to this picture. The Aggressive-Bad group, on the other hand, not only produced more stories with Sex, but distributed the sexual themes through all ten stories. Therefore, the result shown in Table 4-A for two or more stories is in effect what would be obtained by removing Card 13MF from the series. From this it may be concluded that the presence of sexual themes in stories produced in response to pictures not directly suggestive of sexual themes is associated with socially unacceptable overt aggression. The absence of sexual themes in stories produced in response to sexually suggestive pictures is associated with the Passive type of overt behavior.

Table 4-B, C, D, E shows the marked differences between the groups on the tabooed themes, the Aggressive-Bad group always far in the lead, with the Aggressive-Good and Passive groups showing no significant differences between them. Tabooed material in TAT stories is clearly associated with socially unacceptable overt aggressive behavior.

DISCUSSION AND CONCLUSIONS

Responses to a projective test may be thought of as consisting essentially of

TABLE 4
BEHAVIOR-SAMPLE ASPECTS

TAT Variable	Number of Stories	Number of Subjects ^a			Chi Square	Chi Square ^d		
		Agg.-Bad	Agg.-Good	Passive		Agg.-Bad vs. Agg.-Good	Agg.-Bad vs. Passive	Agg.-Good vs. Passive
(A) Sex	1 or more	19	17	6	15.23 ^c	.07	11.09 ^c	7.13 ^b
	none	6	8	19				
	2 or more	14	6	4	11.21 ^b	4.09 ^a	7.99 ^a	.12
	1 or none	11	19	21				
(B) Tabooed Sex	1 or more	7	2	2	5.22			
	none	18	23	23				
	2 or more	7	0	0	14.92 ^c	10.38 ^b	6.50 ^a	.00
	1 or none	18	25	25				
(C) Tabooed Language	1 or more	9	1	3	9.18 ^b	5.16 ^a	3.24	.09
	none	16	24	22				
(D) Tabooed Violence	1 or more	10	0	0	22.41 ^c	8.92 ^b	10.92 ^c	.00
	none	15	25	25				
(E) Combined Taboo (B, C, D)	1 or more	13	3	3	13.81 ^c	6.06 ^a	8.36 ^b	.03
	none	12	22	22				

^a $p < .05$.^b $p < .01$.^c $p < .001$.^d With Yates' correction.^e See footnote b in Table 2.

two kinds of elements: those which are fantasy projections of inner tensions (drive states or conflicts), and those which are samples of the person's behavior. These two elements may be referred to respectively as the "projective" and the "behavior- (or trait-) sample" aspects of the test response. The thematic content of TAT stories carries the largest portion of the projective element, while the so-called formal aspects (*S*'s approach, language, style, perceptual and cognitive use of the stimulus, attitudes toward the task, etc.) represent the behavior- or trait-sampling element.

The results of this study do not support the projective hypothesis that *n* Aggression in TAT fantasy is related to overt aggressiveness as a general disposition or trait. There appeared to be neither a direct nor an inverse relationship between fantasy aggression and overt behavior. As pointed out previously, the direct-relationship hypothesis has found support only in studies dealing with temporary affective states or drive states (3, 6, 12, 16, 17). Thus the TAT, as re-

gards its fantasy element, seems to be more a test of the *S*'s current mood and situational attitude, and possibly also of certain persistent underlying conflicts or needs, rather than a test of overt personality.

The "dynamic hypothesis" is based on the idea that overt behavior is a resultant of a balance of inner forces and that inferences about overt behavior from fantasy content must take into account not only the various drives or "needs" and "presses" represented in fantasy, but also their interaction with each other. Thus the degree of overt aggressiveness may be viewed as a resultant of an aggressive impulse and the anxiety that opposes aggression. The balance of these forces should, according to the dynamic hypothesis, provide the best indicator of the degree of overt aggression.

This hypothesis found some support in the present study, as it has in others (14, 15). An absence of punishment or of defense against aggression anxiety seems to be associated with socially unacceptable overt aggression. The presence of natural death in the stories also seems to serve a defensive purpose and is positively correlated with socially acceptable overt aggressiveness.

The behavior-sample hypothesis requires that evidence of the *S*'s behavioral traits be sought in his test performance. The *S*'s response to the

test situation is regarded as a sample of his mode of responding to other tasks and interpersonal relationships. Examples of this in the clinical use of the TAT (and Rorschach) are the noting of signs of confusion, bizarre verbalizations, perceptual distortions, etc. as signs of schizophrenia; hesitancy, indecision, blocking, etc. as signs of anxiety or obsessiveness; and so on. These are samples of the S's behavior.

It was hypothesized in this study that the Aggressive-Bad Ss would include in their TAT stories defiant, antisocial, or "shocking" elements characteristic of their contempt for school-behavior propriety, and would manifest their lack of inhibition about expressing themselves in socially disapproved ways by giving freer expression to any socially tabooed themes that the pictures might suggest, however remotely. The evidence strongly supported this hypothesis. On each of the "tabooed" themes (Sex, Tabooed Sex, Tabooed Language, Tabooed Violence) the Aggressive-Bad group was far in the lead. This certainly cannot be interpreted as representing gross differences in the strength of the sex drive (*n* Sex) in the three groups. The differences between the groups were consistent in direction on all pictures for all the tabooed variables. This was not the case with any of the other variables in the study. Whereas certain pictures elicited a particular kind of thematic content more than others, the behavior-sample variables, such as Tabooed Language, were elicited about equally by all the pictures.

The obvious conclusion is that the behavior-sample elements of the TAT response, when elicited under the testing conditions described in this study, are much more highly related to overt behavior than the projective or thematic elements. The reason is simple: the formal and trait-sample elements *are* the overt behavior. An analysis of the TAT stories from the behavior-sample point of view reveals more about the S's overt personality than an analysis of the thematic content per se.

It is of course impossible to say with confidence what the results of this study would be if the Ss were girls, or were of a different age group, or were institutionalized delinquents, or mental patients. Also the method of administering the TAT was undoubtedly of significance in the results. The number of sex and tabooed themes would have been much fewer, if they would have been expressed at all, had the test been individually administered. When the Symonds Picture-Story Test (which originally included a picture of two nudes of opposite sex standing face-to-face) was administered individually to 20 high school boys, they gave no responses containing tabooed sex or tabooed language (19).

It should be pointed out that the three experimental groups were not conceived of as lying along a *single* continuum with respect to overt aggression. The main requirement was that the three groups differ markedly with respect to overt aggressive behavior. This was achieved. At least two dimensions would probably be required to account for the behavioral differences among the groups. In retrospect one might hypothesize as the two dimensions: "conformity-nonconformity" and "activity-passivity." The Aggressive-Bad Ss might be characterized as "actively nonconformist," the Aggressive-Good Ss as "actively conformist," the Passive Ss as "passively conformist."

Another point: the selection of the 10 TAT pictures could possibly affect the outcome of such an experiment. The fact that the various TAT cards have different stimulus values is well-known. Certain pictures elicit more of a particular theme than do others. A full presentation of the data of this study (9) clearly reveals that, in addition to this fact, the *proportions* in which certain themes are elicited by various pictures are not the same for different types of Ss. For example, Card 17GF elicited the most fantasy aggression from the Passive group and the least from the Aggressive-Bad group, while Card 20 was just the reverse. The implication would seem to be that the same variable may carry a different weight when elicited by one picture than by another. Aggression elicited by Card 17GF was *negatively* correlated with overt aggressiveness, while aggression elicited by Card 20 was *positively* correlated with overt aggressiveness. Future research might apply an "item analysis" technique to the TAT, treating the individual cards as items and correlating the themes elicited by each one with various personality characteristics. It is not improbable that such an approach to TAT analysis could have greater predictive value than the more global or additive approaches, in which the differential stimulus values of the various pictures are not so systematically taken into account.

Finally, an observation that is made too seldom in studies such as this, is the fact that, even though we were here dealing with groups differing in the extreme, the TAT variables discriminated very poorly for practical purposes. The statistically significant discriminations must be regarded as of only theoretical interest. Using any combination of these statistically significant variables for predictive purposes, even with extreme

groups, would obviously result in a very large percentage of errors, as may be seen from Tables 2, 3, and 4. That the TAT can be used for individual prediction of overt behavior is thus seriously questioned by these results, especially as regards the use of fantasy "needs" and "presses." Somewhat more confidence may be had in behavioral predictions from the TAT when the S's performance is interpreted as a behavior-sample.

SUMMARY

The relationship of Thematic Apperception Test responses to overt behavior was investigated with respect to aggression. The central question was: How is behavioral aggressiveness reflected in the TAT? The major hypotheses were: (a) There is a direct relationship between aggression in TAT fantasy and in overt behavior. (b) There is an inverse relationship between the amount of fantasied punishment, relative to fantasied aggression, and overt aggressive behavior. (c) There is an inverse relationship between the amount of defense against aggression anxiety in the fantasy and overt aggressive behavior. (d) There is a direct relationship between those aspects of the TAT response which may be viewed as behavior-samples and overt behavior in social situations. With respect to socially disapproved overt aggressiveness, the TAT responses would be expected to evince samples of socially disapproved, tabooed, or defiant behavior.

The method of the study was to permit whatever relationships that might exist between the TAT and the behavioral variables the maximum opportunity for manifesting themselves, by comparing relatively homogeneous groups of Ss differing widely in their overt behavior with respect to aggressiveness. From an

all-boy high school three groups of Ss were selected by means of teacher ratings: (a) 25 who were aggressive in socially unacceptable ways (Aggressive-Bad); (b) 22 who were aggressive in socially acceptable ways (Aggressive-Good); (c) 27 who showed no overt aggressiveness (Passive). The Ss wrote stories to 10 TAT cards selected for their tendency to elicit themes of aggression. The stories were scored on the following variables: Aggression (Hero Aggressor, Hero Victim, Indeterminate), Punishment, Defense (against aggression anxiety), Suicide, Natural Death, Sex, Tabooed Sex, Tabooed Violence, and Tabooed Language. Interrater agreement was determined and found to be high (95%).

The "direct relationship" hypothesis was not substantiated. The Aggressive-Bad and Passive Ss had about the same amount of aggression in their TAT fantasies. The Aggressive-Good group had even less fantasy aggression than the Passive group. The groups did not differ in the proportion of aggressive stories in which the Hero was the aggressor or the victim of the aggression. Nor did the groups differ in the amount of mild or socially acceptable forms of aggression in their stories.

The three groups differed at the .01 level of significance in the proportion of aggressive stories which contained neither Punishment nor Defense. The Aggressive-Bad group had a greater proportion of such unmodified or "raw" aggression.

The theme Natural Death was hypothesized as being a defense against aggression anxiety, representing an aggressive wish in disguise. The Aggressive-Good group had significantly more Natural Death in their stories than the other groups.

The groups did not differ significantly in the number of Suicide themes.

The behavior-sample aspects of the TAT stories showed highly significant differences between the groups. All the differences on the following variables were significant at the .01 to .001 level of confidence, with the Aggressive-Bad group far in the lead and the Aggressive-Good and Passive differing from each other hardly at all: Sex, Tabooed Sex, Tabooed Language, and Tabooed Violence. This relationship held up for these variables on all ten TAT pictures; this was not the case with any of the other variables in the study.

The following general conclusions

were drawn: (a) There was very little, if any, relationship between aggression in fantasy and in overt behavior. (a) The absence of themes of punishment and of defenses against aggression anxiety in the TAT was associated with socially unacceptable forms of overt aggression. (c) Aspects of the TAT responses which were regarded as behavior samples were related to overt aggressive behavior at a high level of significance. Ss who habitually acted-out aggressively in ways regarded as taboo in the school setting responded also to the TAT with socially tabooed content and language.

REFERENCES

1. ATKINSON, J. W., & MCCLELLAND, D. C. The projective expression of needs: II. The effect of different intensities of the hunger drive on thematic apperception. *J. exp. Psychol.*, 1948, 38, 643-658.
2. BALKEN, E. R., & MASSERMAN, J. H. The language of fantasy: III. The language of the fantasies of patients with conversion hysteria, anxiety state, and obsessive-compulsive neurosis. *J. Psychol.*, 1940, 10, 75-86.
3. BELLAK, L. The concept of projection: an experimental investigation and study of the concept. *Psychiatry*, 1944, 7, 353-370.
4. CLARK, R. A. The projective measurement of experimentally induced levels of sexual motivation. *J. exp. Psychol.*, 1952, 44, 391-399.
5. COHEN, A. K. *Delinquent boys: The culture of the gang*. Glencoe, Ill.: The Free Press, 1955.
6. FESHBACH, S. The drive-reducing function of fantasy behavior. *J. abnorm. soc. Psychol.*, 1953, 50, 3-11.
7. GLUCK, M. R. The relationship between hostility in the TAT and behavioral hostility. *J. proj. Tech.*, 1955, 19, 21-26.
8. HARTMAN, A. A. An experimental examination of the thematic apperception technique in clinical diagnosis. *Psychol. Monogr.*, 1949, 63, No. 8 (Whole No. 303).
9. JENSEN, A. R. Aggression in fantasy and overt behavior. Unpublished doctoral dissertation, Teachers College, Columbia Univer., 1956.
10. KORNER, ANNELIESE F. Theoretical considerations concerning the scope and limitations of projective techniques. *J. abnorm. soc. Psychol.*, 1950, 45, 619-627.
11. LINDZEY, G. An experimental examination of the scapegoat theory of prejudice. *J. abnorm. soc. Psychol.*, 1950, 45, 296-309.
12. MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
13. MURRAY, H. A. *Thematic Apperception Test manual*. Cambridge: Harvard Univer. Press, 1943.
14. MUSSEN, P. H., & NAYLOR, H. K. The relationship between overt and fantasy aggression. *J. abnorm. soc. Psychol.*, 1954, 49, 235-240.
15. PITTLUCK, PATRICIA. The relation between aggressive fantasy and overt behavior. Unpublished doctoral dissertation, Yale Univer., 1950.
16. SANFORD, R. N. The effects of abstinence from food upon imaginal processes: a preliminary experiment. *J. Psychol.*, 1936, 2, 129-136.
17. SANFORD, R. N. The effects of abstinence from food upon imaginal processes: a further experiment. *J. Psychol.*, 1937, 3, 145-159.
18. SANFORD, R. N., ADKINS, M. M., MILLER, R. B., COBB, E. A., & others. Physique, personality and scholarship: a cooperative study of school children. *Monogr. Soc. Res. Child Developm.*, 1943, 8, No. 1 (Serial No. 34).
19. SYMONDS, P. M. *Adolescent fantasy: an investigation of the picture story method of personality study*. New York: Columbia Univer. Press, 1949.

(Accepted for publication March 11, 1957)

Bureau Ednl. Psy. Research

DAVID HA

MINING COLLEGE

Dated

4-1-58

A Pattern Analysis of Descriptions of "Best" and "Poorest" Mechanics Compared with Factor-Analytic Results¹LOUIS L. MCQUITTY²

Michigan State University

THE particular method of pattern analysis developed in this paper is called *agreement analysis*. It can be introduced by contrasting it with factor analysis. These two methods differ in terms of the assumptions they imply concerning the interrelationships in subject matter they are designed to analyze. Agreement analysis is based on a generalized theory of types. The types are assumed to reflect themselves in many kinds of interrelationships among characteristics—not just linear ones. Consequently, agreement analysis is a tool for determining all of the various ways in which characteristics are presumed to concatenate to yield types. Having this general ability to isolate types in terms of all kinds of interrelationships, it is, of course, capable of isolating them in terms of any restricted kind of interrelationships among characteristics.

Factor analysis, on the other hand,

assumes that characteristics concatenate in terms of linear interrelationships only. It is a tool for analyzing the linear interrelationships among characteristics and for expressing these in terms of factors. Types which may be discovered are limited either to standings on single factors or configurations of standings on several factors.

A more detailed contrast between factor analysis and agreement analysis will help give additional meaning to the latter approach. The most common method of factor analysis (*R*-technique) has as its purpose to take the intercorrelations between a large number of indexes of personality and analyze these to yield a few fundamental personality traits in terms of which the indexes can be classified.

An important consideration in this factor analytic procedure is that the intercorrelations are taken across all subjects of a sample; no allowance is given to the possibility that variables might be interrelated differently in various categories of subjects of a single sample. If variables are interrelated differently among categories of subjects and if the number of subjects representing each category varies considerably, then intercorrelation across all subjects as generally taken in factor analysis would represent a hodgepodge.

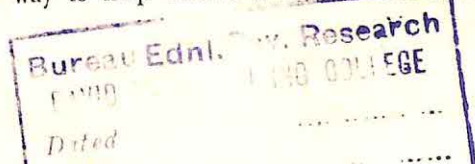
Agreement analysis assumes types and is a method designed to assist in their isolation; it attempts to isolate them by classifying together individuals who give identical responses to large numbers of items of a test.

Substantiation of a need for a method which contrasts with factor analysis is simple. Factor analysis does not prove the existence of factors; it merely clarifies the characteristics which they have if we postulate them.

A limiting characteristic of factor analysis is the fact that it does not include a statistic for rejecting it as not fitting data particularly well. One way to help correct for this defect is to

¹This research was supported in part by the United States Air Force under Contract No. AF 33(038)-25726 monitored by the Commanding Officer, Air Force Personnel and Training Research Center, Lackland Air Force Base, San Antonio, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government. This research was completed while the author was at the University of Illinois.

²Appreciation is expressed to Professor Lyle H. Lanier for both his editorial suggestions and valuable critique which were very helpful in the preparation of this paper.



develop methods based on the assumption of constructs other than factors, and compare results obtained with the two methods.

Agreement analysis postulates one kind of construct (types) which contrasts with factors. Applications of agreement analysis will help us to learn some of the characteristics which its types have. Comparisons of results from factor and agreement analyses will assist us in determining which alternative postulate is more helpful for various purposes.

This paper includes four main sections. The first sections develops the theory and method of agreement analysis and illustrates the method by applying it to descriptions of "best" and "poorest" mechanics. The second section compares the results of the agreement analysis with those of a factor analysis of common data, and interprets the types in the light of both the factor and agreement analyses. The third section critiques the method of agreement analysis. The fourth section summarizes the entire paper.

I. CLASSIFICATION OF DESCRIPTIONS BY AGREEMENT ANALYSIS

As mentioned above, an investigation by agreement analysis is based on a theory of *types*. In the particular theory of the present investigation, the concept of type is treated as an intrinsic construct. Each type is assumed to reflect itself in a pattern of characteristics. Consequently, individuals can be classified into typological categories according to the patterns of characteristics which they exhibit. Each category into which individuals are classified is said to represent a type. The individuals of any one type are presumed to have certain characteristics in common. These characteristics are called a *pattern*. The individuals of any other type are also presumed to exhibit a pattern of characteristics, and their pattern is unique to their type. In brief, the individuals of each type ex-

hibit a pattern of common characteristics which is peculiar to them. For example, Type X might include individuals who exhibit a pattern of dependability, carefulness, cooperativeness, and industriousness; Type Y might include individuals who exhibit the pattern of carefreeness, friendliness, joviality, and slipshodness.

The Linnaean approach (1, pp. 499-492) to the classification of living organisms in the field of biology results in a chart which shows (a) all types of living organisms and (b) the patterns of characteristics peculiar to each type. At the lowest level of classification of living things, each organism is classified into a *species*. Consequently, each species includes a number of organisms. The organisms of each species exhibit a number of common characteristics which are peculiar to them.

At the next level of classification, each species is grouped with other species to form *genera*. Consequently, each genus contains several species. The species of each genus have certain characteristics in common which are peculiar to them. In an analogous fashion, genera are grouped to form *families*, families to form *orders*, and so on through *classes*, *phyla*, and *kingdoms*.

A simplified and incomplete Linnaean chart is shown in Fig. 1 for illustrative purposes. At the bottom level of the chart, we have several individual organisms, then at the next level we have several species, then several genera, etc., up to a class. There is, of course, a pattern of characteristics for each species, genus, family, etc.

Analogously to the simplified Linnaean Chart, we propose to build a classification of certain descriptions of mechanics.

Since so little is known about the types into which descriptions of mechanics may ultimately be classified, we propose for the present to use the Linnaean terms of species, genera, families, etc., for categorizing them.

We have six purposes in building this classification; three of them are methodological, and the other three are substantive. The *methodological* purposes are (a) to illustrate a quantitative method for developing the classification, (b) to

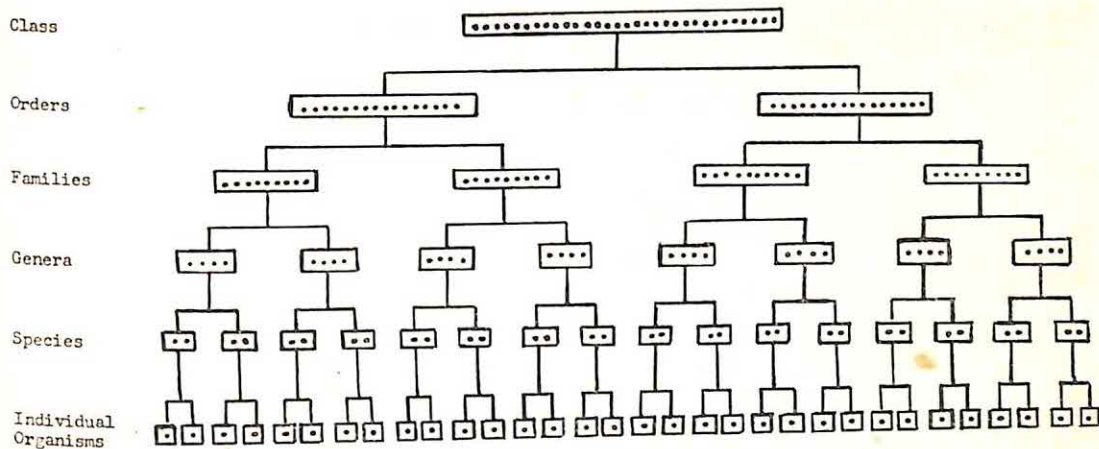


FIG. 1. A simplified and incomplete Linnaean chart.

indicate some of the potentialities of the method in the field of personnel research, and (c) to compare results from this pattern analytic approach with a factor analysis of common data.

The *substantive* purposes are to make a preliminary investigation of the following hypotheses: (a) that the descriptions of mechanics selected as "best" and "poorest" can be classified into types; (b) that the classifications are relatively dependable in the sense that two separate classifications of subjects on the basis of two sets of test items will yield similar results; and (c) that the classifications are valid in the sense that they are related to an external criterion.

The above purposes are, of course, related to a theoretical, psychological position which should be specified unless it otherwise be misunderstood. We are not attempting, initially, to isolate the types which may in some way be said to be fundamental for the generalized understanding of human behavior. For the time being, at least, it is assumed that different types may be significant for different purposes. For example, types which derive from an analysis of symptoms of diseases may be significant for differentiating between mentally and physically ill patients, but other types which derive from an analysis of characteristics of "best" and "poorest" mechanics may be required in differentiating subjects in relation to their qualifications as mechanics. We are attempting to isolate types which are significantly related to a cri-

terion, but we are not arguing that these single criterion types are the most fundamental ones in terms of which to understand behavior in general. On the contrary, we are inclined to the point of view that behavior can be categorized in many different ways and that each way will give somewhat different types.

Even though we are interested initially in types pertinent to single criteria, we are, nevertheless, also oriented ultimately toward the more fundamental scientific problem of deriving a parsimonious, systematic structure for the inclusive representation of behavioral organization. Types which would fulfill this requirement would be minimal in number, simple in description, and of maximal value in predicting behavior.

The single-criterion studies can assist in the solution of the more fundamental problem. The types which derive from the single-criterion studies can be used as the basis for the development of those which are more generally meaningful, and therefore more basic, scientifically. Each single-criterion study will have selected a set of items, significant for the criterion under consideration. All of the items taken jointly from several single-criterion studies and analyzed simultaneously in such a manner as to isolate all the types represented in the data should produce those more comprehensive types which are pertinent to all the several single criteria. Consequently, these types should be relatively fundamental in the sense that they would be generally applicable in the prediction of behavior.

Since there is probably no absolute limit to the number of single criteria which may be used, and since each additional one may add something, it will probably always be possible to improve somewhat on the classification into fundamental types, which derive from this approach.

However, the rate of improvement is probably asymptotic, or at least would become so at some point, so that eventually the amount of improvement which could be added would become very small.

The Method of Classification

We are now prepared to outline and illustrate with data a method of agreement analysis which we have developed for the isolation of types. It will here be outlined as it is applied in a single-criterion study.

Subject and tests. Agreement analysis was applied to data which had been gathered in an earlier factor-analytic study by McQuitty, Wrigley, and Gaier (3). As pointed out in this earlier paper, an inventory of 200 items was developed using the language which mechanics used in describing other mechanics whom they had selected as representative of either "best," "poorest," or "average" mechanics. The inventory for describing mechanics was administered to 428 supervisors, who selected a "best," a "poorest," or an "average" mechanic and described him by replying either "yes" or "no" to each of the 200 items of the inventory. In the present investigation, we used the results from supervisors who described either a "best" or a "poorest" mechanic; we did not use the results from those who described an "average" mechanic.

In the first classification, we analyzed results from the first 120 items. In the second classification, we analyzed results from the remaining 80 items. The 200 items of the inventory were listed in random order. The initial classification was based on the first 120 items because the factor analysis mentioned above was in terms of these items, and we wished to compare results from the two different kinds of analyses.

We shall describe the first classification in

detail; the second one will be clear by analogy. For each supervisor in the first classification, we had 120 responses which described a mechanic selected by him as either a "best" or a "poorest" mechanic. Each response is a description of a mechanic. Each 120 responses by any one supervisor is a pattern of descriptions. Each pattern of descriptions pertains to a particular mechanic, either a "best" one or a "poorest" one.

There were 112 patterns pertaining to "best" mechanics and 92 pertaining to "poorest" mechanics, giving a total of 204 patterns. Using random numbers, we arbitrarily reduced the 112 patterns for "best" mechanics to 108 to give a grand total of patterns for "best" and "poorest" mechanics, combined, of 200 (for convenience of machine computations).

Again using random numbers, we divided the 108 patterns for "best" mechanics into equal-sized experimental and cross-validated samples. We did likewise for the 92 patterns descriptive of "poorest" mechanics. The two experimental samples, one for "best" and the other for "poorest" mechanics, were used to build our classification of patterns into types. The cross-validated samples were used to determine how well the types held up on other samples of the same universe.

As the patterns of descriptions are classified, successively, into more and more inclusive types, some of the components of the patterns are dropped. In other words, the initial patterns with which we start are reduced in their numbers of descriptions as the classification proceeds into more and more inclusive types. In order to differentiate these several "levels" of classification, we shall speak of *individual patterns* when classifying the initial ones into *species*; then of *species patterns* when we are classifying into *genera*; and of *genus patterns* when we are classifying into *families*; and so on.

Details of the method. The method of analysis here used is an abbreviated version of a complete method of pattern analysis, developed by the author and called agreement analysis (6, 7). The present version differs from the complete method in these respects: (a) it omits a calculation which corrects agreement scores for chance fluctuations; (b) it applies a somewhat less exacting, but more rapid, technique for classifying patterns into the categories which they resemble most; and (c) it limits each category to *not more than two patterns from the*

next lower level of classification; it is therefore a binary version.

This abbreviated version of agreement analysis is justified here because the purpose is not to isolate the most fundamental types but rather to study patterns of responses in relation to an external criterion. This version allows the investigator to isolate rather rapidly a great number of reasonably reliable patterns and to study them as predictors.

Our abbreviated method of binary agreement analysis involves a novel treatment of the agreement score developed by Zubin (10). The agreement score is computed for pairs of individual patterns. The agreement score for any two individual patterns is the number of identical descriptions contained in the two patterns. Suppose that the two individual patterns A and B both contain "yes" for the first five items of the inventory, disagree on the sixth item (one pattern having a "yes" and the other pattern a "no"), and that both patterns contain "no" for the next four items. On these 10 items, the two patterns, A and B, would have an agreement score of nine, because their answers would agree on nine of the 10 items.

Each individual pattern was paired with every other individual pattern, and an agreement score was determined for each pair of individual patterns. These agreement scores are illustrated in Table 1 which shows 10 of the individual patterns descriptive of "best" mechanics. The patterns are labeled A through J and are so designated in the top row and left hand column of the table.

The number, 102, in the intersection of row

A and column B of the table is the agreement score of the pair of individual patterns A and B. This same score also occurs in the intersection of row B and Column A. Other entries in Table 1 are to be interpreted in an analogous fashion. Entries in the diagonal cells are omitted, since they would report the agreement score of each individual pattern with itself and are not needed in the analysis.

In the present analysis, Table 1 was expanded to include the agreement scores for all 54 individual patterns descriptive of "best" mechanics in the experimental sample, i.e., the sample on which the classification of the descriptions of "best" mechanics was developed. Table 1 is shown as an abbreviated illustration of this larger matrix of agreement scores.

The largest agreement score in the 54×54 matrix was selected. Of all the individual patterns represented in the matrix, the two with the largest agreement score have more answers in common than any other two. These two patterns constitute the first species. They and all agreement scores for either of them are withdrawn from the matrix. The pair of patterns with the highest agreement score in the reduced matrix is next selected. This constitutes species number two. The remaining species are selected in an analogous fashion.

In the case of a tie in agreement scores, such as between the individual patterns of pair K and L, on the one hand, with M and N, on the other, it is immaterial as to which species is withdrawn first. However, in the case of a tie such as between X and Y, on the one hand, with X and Z, on the other, it is material. In this latter kind of tie with one pattern common to two pairs, we have followed the practice of classifying X with Y rather than Z only if the next highest agreement score in which Y occurs is lower than the next highest in which Z occurs. By selecting

TABLE 1
ILLUSTRATIVE SAMPLE OF "BEST" MECHANICS' AGREEMENT SCORES BETWEEN INDIVIDUAL PATTERNS

	A	B	C	D	E	F	G	H	I	J
A		102	90	93	93	99	93	91	92	99
B	102		94	93	93	97	95	93	94	99
C	90	94		101	101	95	97	97	98	100
D	93	93	101		100	100	98	92	99	98
E	93	93	101	100		102	96	92	95	94
F	99	97	95	100	102		92	90	97	96
G	93	95	97	98	96	92		100	97	100
H	91	93	97	92	92	90	100		103	100
I	92	94	98	99	95	97	97	103		105
J	99	99	101	98	94	96	100	100	105	

NOTE.—The table is to be read as follows: The agreement score between individual patterns A and B is 102; between individual patterns A and C is 90; etc. For definition of "individual pattern" and "agreement score," see text.

Y instead of Z, the agreement score which will be used to classify Z is larger than would be the one to classify Y if we had reversed our choice. If the second highest agreement scores are also tied, reference can be made to the highest ones which are not tied. These can then be used as a basis for a decision.

We have now outlined how to classify all 54 individual patterns into species, *each species containing two members, and only two*. In those studies where there is one individual pattern left over, it is classified into an additional species containing only one case.

The classification of individual patterns into species started with a pattern of answers for each mechanic described. The pattern of answers is the endorsements on the particular inventory used to describe the mechanic. The classification into species brings together those pairs of individual patterns which are highly similar, as determined by agreement scores. We can use this same approach for classifying species into genera, provided we can obtain a pattern of answers for each species.

A species' pattern of answers is those endorsements which are common to both individual patterns of the species. Suppose, for example, that both individual patterns X and Y of a species have "yes" for item 1 and "no" for item 2. Their species pattern for the first two items would be "yes"-*"no."* Suppose further that individual pattern X has "yes" for item 3, while individual pattern Y has "no" for this item, and they both have "yes" for item 4. Their species pattern for the first four items would be "yes"-*"no"*-blank-*"yes."* The blank denotes the fact that the two individual patterns disagree on item 3. This item is dropped from consideration in the further classification of this species. In other species where both individual patterns have the same answer to the item, it is retained.

In the manner just described, a pattern of answers is determined for each species which contains two individual patterns. These are called species patterns. In those studies where the analysis started with an odd number of cases and ends in the last species containing but one individual pattern, the pattern of answers for this species is the same as those of the one individual pattern which the species includes.

Since we now have a pattern of answers for each species, we can use the same technique for classifying species patterns into genera as we originally used for classifying individual patterns into species. Further, we can use the same plan for determining the pattern

of answers for each genus as we used for determining the pattern of answers for each species. The system is entirely repetitive at each successive level; we merely continue it for classifying species patterns into genera, genus patterns into families, etc., until the classification is completed.

In an analogous fashion as outlined above, we classified the patterns descriptive of "poorest" mechanics. This classification had to be discontinued at the class level, where there were only two classes, because the patterns for classes contained common responses to no item.

Unique features of the method. The method of analysis contains a number of unique features. It is essential to appreciate these if the implications of the method are to be understood and if results from it are to be properly interpreted. These unique features can be introduced by showing how the method is designed to capitalize on the predictive possibilities of responses treated in combinations.

Meehl (9) has illustrated that it is theoretically possible for responses treated in pairs to have relationships to a criterion even though the items treated separately have zero relationship to a criterion. This potentiality of responses treated in pairs would appear to make it highly desirable to analyze responses in combinations of two. However, 100 two-answer items taken two at a time would yield 4,950 pairs of items and 19,800 response pairs for analysis. To analyze this many combinations would be an unreasonably laborious task. Besides, to analyze in combinations of two is not an entirely satisfactory solution. For, if pairs have potentialities over and above separate items, then by the same logic triads have potentialities over and above pairs, and likewise any order of a

combination has potentialities over and above any one of lower order. We are forced then to the conclusion that it is essential to treat items of a test in relatively large combinations if we are to investigate thoroughly the predictive potentialities of the test.

To analyze all items of a test in combination is not nearly as laborious as it would at first seem. Even though there are many combinations of responses theoretically possible, only a very few of them occur empirically, and those which do not occur need not concern us.

In those patterns which do occur, some of the responses are relatively undependable and may be disregarded. A problem is to determine which responses are relatively undependable. In our approach to this problem, we assume that a response may be completely dependable for some subjects and completely undependable for other subjects.

This assignment of differential, all or none, dependability of responses *in relation to the subjects who give them* is novel in test analysis. The usual approach is to assign a single index of reliability for each item as computed from the average behavior with respect to it, even though some subjects may always give the same response to an item while others continually change in their responses.

Our approach, on the other hand, assigns a perfect reliability to an item for those subjects whose responses to it are consistent with their responses to other items. For the same item, our approach assigns a zero reliability for those subjects whose responses to this item are inconsistent with their responses to other items.

In an effort to determine which responses are consistent with other responses, we classify subjects together as already explained in our method

of agreement analysis. In this approach, some items are irrelevant (undependable) in the sense that they do not classify a pattern in the same fashion as the majority of the responses in a pattern. As an illustration, suppose that a test of 100 items is administered to 100 subjects, and that subjects A and B agree in the answers which they give to 99 of the 100 items. Suppose further that the next highest agreement score that either of these subjects has with any other subject is 80. We would conclude that A and B should be classified together in terms of their patterns of responses to the items of the test, and that their responses to the one item on which they disagree is irrelevant to their classification. Their classification would be achieved in terms of their predominant patterns of responses, where these patterns are defined as the ones which include the greatest possible number of responses and still yield a classification.

Consequently, any response which does not conform to the predominant pattern of responses is presumed to be irrelevant. It is assumed to vary independently of the responses in the predominant pattern. In computing a correction for agreement scores in the complete method of analysis (6, 7), irrelevant responses are assumed to vary by chance with respect to the predominant patterns.

A response which is irrelevant in the classification of one subject may, however, be relevant in the classification of another subject. Suppose that the one item on which subjects A and B of the above example disagree is item 2; their responses to this item are irrelevant in classifying them together. Suppose further that C and D are classified together because they agree on 99 (of 100 items) including item 2 but excluding item 3. In this case item 3 rather than item 2 would be designated as irrelevant; item 2 would be relevant for subjects C and D but not for A and B.

As the classification by agreement analysis proceeds from the species levels to higher ones, the criterion for judging responses to be relevant usually becomes more stringent. In order for a response to an item to be relevant in the first classification, a response to it has only to be given by both of two subjects who classify together. As the classification proceeds, more and more individuals are classified into fewer but larger categories and all of the subjects of any one category have to agree on a response in order for it to be relevant to their classification.

Since irrelevant responses are assumed to be independent of the predominant patterns which determine the classification, they are presumed to occur by chance with respect to the predominant ones. Consequently, irrelevant responses can be expected to yield both agree-

ments and disagreements between predominant patterns. We can spot those which yield disagreements as shown above. From these, we can estimate the number of irrelevant responses which yield agreements in any classification. For example, if subjects A and B who agreed on 99 out of 100 items of a test were responding to items which have only two possible answers, we would expect that their agreement on irrelevant responses would equal their disagreement on irrelevant responses. Since they disagreed on one irrelevant item, we would estimate that they agreed by chance on one irrelevant item and we would further estimate their true agreement score to be 98 rather than 99. This example illustrates the logic on which our corrected agreement score is based for the complete version of agreement analysis (6, 7).

It is helpful to emphasize the assumption on which the correction is based in order that agreement analysis can be properly understood. The correction does not assume that each answer alternative is equally apt to occur, as many colleagues have originally thought when they first examined the method. It assumes instead that subjects are equally apt to give any one of the possible answer alternatives to those particular items which are irrelevant for their classification into predominant patterns.

Results from the Classification

A portion of the results from classifying descriptions of "best" mechanics is shown in Table 2; space does not permit showing all of them. In the table, the lowest level of classification represents the individual patterns; the next level represents species; and then a genus is shown. This table corresponds to only a small portion of the chart shown as our model in Fig. 1. It corresponds to any section which includes one genus, two species, and four individual patterns.

Table 2 shows certain detailed information not shown in Fig. 1. The capital letters of Table 2 are codes for patterns descriptive of "best" mechanics. Thus the entry on the left side of the table at the species level means that this species includes individual patterns K and L. Under this species is a twice underlined number, 106; this is the number of items on which individual patterns K and L agree. It is their agreement score, out of a total of 120 times. At the genus level, we find the codes K-L and H-I.

TABLE 2

AN ILLUSTRATIVE CLASSIFICATION OF DESCRIPTIONS OF "BEST" MECHANICS

Genus: K-L-H-I			
Agreement score: <u>86</u>			
Species: K-L Agreement score: <u>106</u>		Species: H-I Agreement score: <u>103</u>	
"No" Ans.: <u>4</u> , <u>8</u> , <u>32</u> , <u>33</u> , <u>40</u> , <u>61</u> , <u>70</u> , <u>75</u> , <u>89</u> , <u>103</u> , <u>112</u>		"No" Ans.: <u>5</u> , <u>15</u> , <u>52</u> , <u>71</u> , <u>80</u> , <u>86</u> , <u>116</u>	
"Yes" Ans.: <u>2</u> , <u>5</u> , <u>26</u> , <u>31</u> , <u>56</u> , <u>77</u> , <u>80</u> , <u>111</u> , <u>116</u>		"Yes" Ans.: <u>4</u> , <u>9</u> , <u>40</u> , <u>49</u> , <u>50</u> , <u>59</u> , <u>66</u> , <u>87</u> , <u>96</u> , <u>112</u>	
Individual Pattern: K	Individual Pattern: L	Individual Pattern: H	Individual Pattern: I
"Yes" Ans.: <u>9</u> , <u>15</u> , <u>49</u> , <u>52</u> , <u>71</u> , <u>86</u> , <u>87</u> , <u>96</u> , <u>100</u>	"Yes" Ans.: <u>50</u> , <u>55</u> , <u>59</u> , <u>66</u> , <u>81</u>	"Yes" Ans.: <u>2</u> , <u>32</u> , <u>33</u> , <u>61</u> , <u>70</u> , <u>77</u> , <u>100</u> , <u>111</u>	"Yes" Ans.: <u>8</u> , <u>26</u> , <u>31</u> , <u>55</u> , <u>56</u> , <u>75</u> , <u>81</u> , <u>89</u> , <u>103</u>
"No" Ans.: <u>50</u> , <u>55</u> , <u>59</u> , <u>66</u> , <u>81</u>	"No" Ans.: <u>9</u> , <u>15</u> , <u>49</u> , <u>52</u> , <u>71</u> , <u>86</u> , <u>87</u> , <u>96</u> , <u>100</u>	"No" Ans.: <u>8</u> , <u>26</u> , <u>31</u> , <u>55</u> , <u>56</u> , <u>75</u> , <u>81</u> , <u>89</u> , <u>103</u>	"No" Ans.: <u>2</u> , <u>32</u> , <u>33</u> , <u>61</u> , <u>70</u> , <u>77</u> , <u>100</u> , <u>111</u>

Capital letter = codes for patterns.

Double underlined numbers = agreement scores.

Other numbers = codes for items. The code numbers listed immediately below each of the two species, for example, represent the items for which both individual patterns of the species have the same answer, but for which at least one individual pattern of the other species has the opposite answer. In cases where both individual patterns of the other species have the opposite answer, the code numbers are underlined. Only those items are listed at each level which are dropped at the next higher classification as irrelevant. The two individual patterns of species K-L disagree on 11 items. These items are listed under both individual patterns K and L, but with opposite answers in the two places; the individual patterns of the species agree on the other 106 items of the inventory.

These letters mean that this genus includes species patterns K-L and H-I, which were derived from individual patterns K, L, H, and I. They agree on 86 items as shown by the twice underlined number, 86, listed under the code letters for the genus. Other code letters and other twice underlined numbers of Table 2 are interpreted in an analogous fashion.

The species are arranged from left to right according to the size of agreement scores, with the largest ones on the left. In other words, as we move from left to right across charts such as the one of Table 2, the classifications generally become less dependable, simply because we thus generally obtain successively smaller agreement scores, and they are successively less dependable.

In a complete chart of results (which could be prepared from the analysis), all of the items and responses to them would be listed in such a manner as to show both the exact answers contained in any pattern, and which items differentiate various individual patterns, species, genera, etc., one from another and to what extent. These points can be illustrated by reference to the incomplete chart of Table 2. Consider the items listed immediately below either of two species which classify in a genus. The items below either of these two species are the ones for which each individual pattern of the species has the same answer, and for which at least one individual pattern of the other species has the opposite answer; sometimes both individual patterns of the other species have the opposite answer. In this latter case the code numbers for the items concerned are underlined. The items are classified according to whether the agreement is in terms of a "yes" answer or a "no" answer. For example, both individual patterns of species K-L have a "yes" on the list of items beginning with 2, 5, 26, etc.; and both individual patterns of species H-I have a "no" on the list of items beginning with 5, 15, 52, etc. Items 5, 80 and 116 are common to these two groups, as indicated by the single underlines. For these two species, the latter three items are perfectly positively correlated with each other and perfectly negatively correlated with the other underlined items of the species, i.e., items for which species K-L have "no" answers and species H-I have "yes" answers.

A complete chart would show the entire description contained in each individual pattern, each species pattern, each genus pattern, etc. This can be illustrated by reference to the incomplete chart of Table 2. Suppose that we want to know the composition of individual pattern L. The answers to the items listed immediately below it give only a few of the answers which it includes; some of the others are the answers to the items listed for the species containing the letter L, namely species K-L. Still others are

the answers of the items for genus K-L-H-I. These latter answers are not shown in the incomplete chart of Table 2 because of lack of space. Analogously, as would be shown on a complete chart, all of the additional answers of individual pattern L, would be listed for the family, order, class, and phyla which contain the letter L. The classification was completed at the phylum level because all classes came together in a single phylum. The description of any other initial pattern or any species, genus, family, etc., can be read in an analogous fashion from a complete chart. We see then that there is a very unusual and important feature of a complete chart. Even though it is the terminal table of a method which has analyzed data in order to facilitate interpretation of it, *no piece of information is lost; the response of each subject to each item is shown.*

The above described feature of retaining every piece of information in the summary table (derived from the analysis) is not realized without creating a problem. One purpose of scientific inquiry is to abstract that material which leads to generalizations. If all information is retained, no abstracting has occurred. However, abstracting can occur at various levels. In agreement analysis, each successive level of classification represents an increase in the degree of abstraction. Consequently, the question arises as to which level yields the most meaningful results. From a purely practical point of view, we can argue in favor of that level which yields the closest relationship with an external criterion. However, this principle would appear to be a rather tenuous one on which to base fundamental, scientific decisions. A better scientific basis can be suggested if the problem is considered in the light of the theory out of which agreement analysis developed. This theory postulated real types. If real types exist, they may be either concise and discrete or more amorphous in nature. If they are discrete, then they should reveal themselves most cogently at some one level of classification. *This would probably be the level of classification at which the agreement scores have the highest statistical significance.* Consequently, statistical theory of significance could be applied in an effort to determine the level of classification in which real types are portrayed. Some statistics designed to evaluate the significance of agreement scores have been reported elsewhere (4).

On the other hand, suppose that types exist only in a relatively amorphous sense; those at any one level are ill-defined and fade into those at the next higher level. Under these conditions, there might be no very fundamental basis for stating that the types at any one level are more basic than those at any other level. *The level of primary concern would then have to depend on*

some utilitarian consideration, such as the degree of relation to an external criterion.

There is another unusual feature of agreement analysis. Since the analysis is based on combinations of answers, and since no combination whatsoever is excluded by the method, the method automatically handles whatever kinds of interrelationships (linear, curvilinear, and disjunctive) are involved in the data, without in any way restricting them. The method also allows any specific answer to be given various meanings, or interpretations, depending on the combinations or patterns of other answers with which it occurs.

Reliability of the method. In order to investigate the reliability of the method of classifying subjects into categories, the 100 subjects of the experimental sample who were originally classified on the basis of the first 120 items of the inventory were again classified on the basis of the last 80 items of the inventory. In the original analysis, the 54 descriptions of "best" mechanics and the 46 descriptions of "poorest" ones were classified separately. This same approach was followed here in classifying the subjects on the basis of 80 items, thus resulting in two independent studies of reliability, one on descriptions of "best" mechanics and the other on descriptions of "poorest" mechanics.

Each analysis was continued until every individual was classified into one of two large categories at the top level of classification. However, we did not wish to assess reliability at this highest level exclusively, because as already pointed out we planned tentatively to study later the validity at all levels of classification and we wished to know whether or not the classification was sufficiently dependable to justify validity studies. For this latter purpose, it did not appear necessary to study reliability at all levels, if we could gain an overview from selected portions. Consequently, for the purpose of comparing the classifications on the basis of 80 items with the earlier ones on the basis of 120 items, we selected categories which were intermediate in the analysis. The categories were selected so that every subject was classified once and only once for the 80-item study, and likewise for the 120-item study. In the case of "best" mechanics, we selected for both studies the first class (two orders) of 32 sub-

jects resulting from the analyses and the two remaining orders of 16 and 6 subjects. In the case of "poorest" mechanics, we selected the first order (two families) of 16 subjects and the four remaining families of 8, 8, 6, and 8 subjects. The categories of "best" mechanics were ordered in each analysis according to the number of test items on which the patterns of each category agreed. In the 120-item analysis of "best" mechanics, the 32-subject class agreed on 43 items, the 16-subject order on 31 items and the 6-subject order on six items. These categories were designated 1, 2, and 3 respectively. In the 80-item analysis, the 32-subject class agreed on 36 items, the 16-subject order on 24 items and the 6-subject order on 13 items. These categories were designated *a*, *b*, and *c* respectively.

To the extent that categories *a*, *b* and *c* have common members with categories 1, 2 and 3, respectively, the method of analysis is reliable across items. The correspondence between these two sets of categories is shown in Table 3. Thirty-

TABLE 3
ASSOCIATION BETWEEN TWO AGREEMENT
ANALYSES OF COMMON SUBJECTS
("BEST" MECHANICS)

Categories From Last 80 Items	Categories From First 120 Items		
	3	2	1
<i>a</i>	1	8	23
<i>b</i>	1	7	8
<i>c</i>	4	1	1

four of the 54 cases, 63 per cent, have identical classification in the two analyses. Combining categories 3 with 2 and *b* with *c* to obtain a 2×2 table (with larger expected frequencies), we computed the chi square and found it to be 5.12, significant at about 0.02. (Chi squares of 5.41 and 3.84 are required for $p = 0.02$ and $p = 0.05$ respectively.)

The analogous results for the 5 categories of the "poorest" mechanics are shown in Table 4.

TABLE 4
ASSOCIATION BETWEEN TWO AGREEMENT
ANALYSES OF COMMON SUBJECTS
("POOREST" MECHANICS)

Categories From Last 80 Items	Categories From First 120 Items				
	8	7	6	5	4
<i>d</i>					8
<i>e</i>			1	4	3
<i>f</i>	5		5	2	4
<i>g</i>	1	5			
<i>h</i>	2	1	2	2	1

Twenty-four of the 46 cases, 52 per cent, have identical classifications in the two analyses of "poorest" mechanics. Combining categories 8, 7, and 6; 5 with 4; f, g, and h; and d with e to obtain a 2×2 table (with larger expected frequencies) we computed the chi square and found it to be 16.81, significant at better than 0.001. (The chi square for $p = 0.001$ is 10.83.) These two sets of results show that *binary agreement analysis gives reasonably dependable results even when applied to relatively homogeneous groups such as "best" or "poorest" mechanics.*

Validity of the method. The two classification schemata from the original analysis based on 120 items (one for patterns descriptive of 54 "best" and the other for patterns descriptive of 46 "poorest" mechanics) were investigated as tools for differentiating between cross-validated samples. In order to simplify a discussion of the method used, suppose we have an individual pattern X, which is not one of the experimental patterns, and suppose its categorization as between "best" and "poorest" is kept a secret until we classify it in terms of the two schemata. We then check to see if the classification is correct. In order to classify pattern X in terms of the two schemata, we first obtain an agreement score for it with *each* of the individual patterns of the experimental samples. On the basis of these agreement scores, we classify individual pattern X in one of three ways, namely, (a) as a pattern descriptive of a "best" mechanic, if its highest agreement score is with an experimental "best" individual pattern, (b) as a pattern descriptive of a "poorest" mechanic, if its highest score is with an experimental "poorest" individual pattern, or (c) as

indeterminate, if there is a tie and individual pattern X has its highest agreement score with both a "best" and a "poorest" experimental, individual pattern. The classification schemata are valid for individual pattern X if they classify it correctly.

However, in the above procedure, we have not yet exhausted all of the potentialities of the classification schemata. In fact, we have hardly used the schemata at all. We have scored individual pattern X in terms of individual patterns only. We can also score it in terms of the patterns of species, genera, families, etc., for "best" and "poorest" experimental data and classify it as either "best" or "poorest" at each of these levels. A problem is to discover which level gives results in closest agreement with the criterion classification. This problem is investigated by classifying many cross-validated, individual patterns as "best" or "poorest" at each level and analyzing the results to determine which level gives the highest percentage of correct classifications. We did this for the individual patterns of our cross-validated samples.

The results are shown in Table 5, which reports the percentages of correct classifications for cross-validated, individual patterns descriptive of "best" and "poorest" mechanics, shown separately, and then also combined, with each category weighted equally. For example, at the species level, 100 per cent of the individual patterns for "best"

TABLE 5
PERCENTAGES OF CORRECT CLASSIFICATION ON THE CROSS-VALIDATION SAMPLE (120 ITEMS)

Classification	Individuals	Species	Genera	Families	Orders	Classes	Phyla
"Best"	96	100	100	100	100	100	100
"Poorest"	89	83	76	70	70	33	13
Combined	93	92	88	85	85	67	57

mechanics and 83 per cent of those for "poorest" mechanics were correctly classified. A mean of these two, namely 92, is the average percentage of the combined groups correctly classified. The maximum percentage of correct classifications for the individual patterns descriptive of "poorest" mechanics is 89 and occurs at the individual level. For those descriptive of "best" mechanics, it is 100, for all levels above that of the individual patterns. For "best" and "poorest" combined, and equally weighted as groups, the maximum percentage of correct classifications is 93 and occurs at the individual pattern level.

Multiple types of "best" and "poorest" descriptions. There is one very interesting characteristic of the cross-validated percentages. The largest percentage for the individual patterns descriptive of "poorest" mechanics is at the *individual-pattern* level, with 89 per cent correctly classified; the percentages then show a decrease down to only 33 per cent correctly classified at the *class* level. In other words, as descriptions of both "best" and "poorest" mechanics are classified separately into more inclusive categories, the items in terms of which the "best" and "poorest" descriptions differ drop out of the analysis. These items which differentiate as between "best" and "poorest" descriptions also differentiate within both the "best" and "poorest" groups and consequently yield many types of both "best" and "poorest" descriptions. On the other hand, the items in terms of which many "best" descriptions agree have identical or similar responses for descriptions of "poorest" mechanics; "poorest" descriptions do not differ from "best" ones primarily in those characteristics which are common to "best" descriptions, but rather in those which are relatively

heterogeneous for even "best" descriptions. Our analysis revealed no *single* ideal type of "best" mechanic which differs from an all-inclusive prototype of "poorest" mechanic; *multi-types* seem to be required if maximum differentiation is to be obtained between categories of "best" and "poorest" mechanics.

Hypotheses analogous to the ones just outlined have derived from a similar study by the writer which compared mental hospital patients with community subjects (5). Mental hospital patients do not seem to differ from community subjects on those characteristics which are common to almost all community subjects. Instead, they appear to differ more on those characteristics which are peculiar to species and genera of community subjects. Multi-types of both patients and community subjects appear to be required in order to achieve maximum differentiation. These hypotheses might prove helpful in areas which have frustrated quantitative efforts to differentiate categories of individuals satisfactorily.

Improvement through item selection. The fact that the percentage of correct classifications decreases as we move from species to higher level categories suggests the hypothesis that items which are relevant to the greatest number of species are the most effective in obtaining correct classifications. However, some items are relevant for both species and higher-level categories. In fact, in order to be relevant for a higher-level category, an item must be relevant for all species classified in it. But items which are relevant at both species and higher levels would fail to yield the species. In order to differentiate between species, we must have some items which are relevant *at the species level only*. In attempting to select a reduced number of items which would be equally as effective as the original 120 items, we selected 26 items with the following characteristics: (a) each item selected must be relevant for at least 82 per cent of the species isolated; (b) some of the items selected

must be relevant primarily at the species level only; (c) others must be relevant at higher levels as well. As a consequence, the items selected varied in relevancy at the order level from those which were relevant for only 16 per cent of the orders up to those which were relevant for 71 per cent of the orders.

Using these 26 items, in lieu of the original 120 items, we repeated both the experimental and cross-validated studies. As a control group of items against which to compare the above selected sample, we chose 26 items at random from the 120 original ones and also repeated the studies on them.

The percentages of correct classifications derived from 26 *especially selected* items on the cross-validated samples is shown in Table 6. It is analogous to

TABLE 6
PERCENTAGES OF CORRECT CLASSIFICATIONS
FOR CROSS-VALIDATION SAMPLES (26
ITEMS OF KNOWN RELEVANCIES)

Classification	Individual	Species	Genera	Families	Orders	Classes
"Best"	98	100	100	100	100	100
"Poorest"	89	89	85	72	63	20
Combined	94	95	93	86	82	60

Table 5, which was based upon 120 items. For combined samples ("best" and "poorest") of the cross-validated subjects, the 26 items give higher percentages of correct classification than the 120 items at the individual, species, genus, and family levels; at the order and class levels, the 120 items give the higher percentages. The highest percentage of correct classifications, namely 95, is obtained at the species level for the 26 items.

The percentage of correct classification derived from the 26 *randomly selected* items for the cross-validated subjects are shown in Table 7. This table is anal-

TABLE 7
PERCENTAGES OF CORRECT CLASSIFICATION,
FOR CROSS-VALIDATION SAMPLES,
26 RANDOM ITEMS

Classification	Individual	Species	Genera	Families	Orders	Classes
"Best"	100	100	100	100	100	100
"Poorest"	78	76	78	76	72	48
Combined	89	88	89	88	86	74

ogous to Table 5, based upon the 120-item study, and to Table 6, based upon the 26 especially selected items.

The 26 especially selected items gave the two highest percentages of correct classification, 94 and 95, and tied with the 120 items for the third highest, 93. In addition to this tie, the 120-item test gave the fourth highest percentage, 92. The 26 random items gave the fifth highest, 90, and tied with the 120-item test on the sixth highest, 88.

Since the purpose of selecting items was to obtain those most effective at the species level, we computed the significance of the difference in percentage of correct classifications at this level between the 26 especially selected items and the 26 items chosen at random. Table 8 shows the

TABLE 8
ACCURACY OF CLASSIFYING SUBJECTS WITH
RANDOM VS. SELECTED ITEMS

26 Random Items	26 Items of Known Relevancies	
	Incorrect	Correct
Correct	0	88
Incorrect	5	7

data for this computation, assuming a correlation between the percentages. However, this table does not meet the minimal requirements for a computation of this kind; one cell has a zero entry and the cell diagonally below it has only seven as an entry (2, p. 80). A test which fails to assume a correlation can, however, be applied to this data provided the result is interpreted with the understanding that it is an underestimate of the significance of the difference to the extent that the percentages are correlated, as they would normally be in a situation of this kind. A test of this kind (2, p. 76) was applied. It produced

a critical ratio of 1.88, which is significant at the 0.03 level in a single-tail interpretation, as is here appropriate.

The empirical results reported above support the hypothesis that items which have high relevancies for species and vary in their relevancies for higher-level types give higher percentages of correct classifications than an equal number of randomly selected items.

In cases where an agreement analysis fails to give a satisfactory percentage of correct classifica-

tions, there is sometimes a possibility of improving on the results. This possibility derives from the fact that one application of agreement analysis to all of the items of a test isolates only the predominant types. The responses which do not fit these types are rejected at the various levels of classification. The pattern analysis could be repeated on the rejected responses. These would yield species not isolated from the analysis of the test as a whole. They might give correct classification missed by the other items. In the present study, we had relatively few incorrect classifications and relatively few responses rejected at the species level. As a consequence, we did not try a reclassification on the basis of rejected items;

TABLE 9
43 ITEMS DESCRIPTIVE OF THE FIRST CLASS (2 ORDERS) OF "BEST" MECHANICS ($N = 32$)

Item No.	Response	Item
1	Y	He can get along with people pretty well.
11	Y	When he does a job, you know it will be done right.
13	Y	He answers questions as best he can.
17	Y	He knows how to use tools.
18	Y	His personal traits are O.K.
39	Y	You can take his word for what he says he's done.
44	Y	He deserves a promotion.
47	Y	You don't have to worry about telling him what to do all the time.
57	Y	He sees decision through to an end.
60	Y	He uses the right tools when he can get them.
67	Y	He is very glad to help somebody else.
68	Y	He makes sure he does a good job.
72	Y	He tries to find better ways of doing things.
73	Y	His work is nearly 100% correct.
78	Y	The pilots all seem to get along with him pretty well.
85	Y	What he doesn't know he can find out for you one way or another.
90	Y	Everyone enjoys working with him.
91	Y	He knows his stuff.
97	Y	If you leave him to do a job, you can always be sure he will get the job done.
98	Y	What he knows, he never forgets.
101	Y	He usually remembers things which are explained to him.
105	Y	He can show you how to do the job right.
106	Y	His ambition will pay off.
107	Y	He is good at working on the plane.
118	Y	He gives good cooperation.
119	Y	He seems to take pride in his work.
7	N	He is more often wrong than right in his decisions.
10	N	He works in a sloppy way.
21	N	He is afraid to make a decision one way or the other.
27	N	Most guys with that much experience know a lot more than he does.
28	N	He just stands around until you're done before offering to help.
30	N	He doesn't have any sense of responsibility.
35	N	He's just careless.
38	N	He isn't a very careful worker.
41	N	He achieves his aim in the wrong way.
46	N	He's always getting messed up in money deals.
69	N	He is kind of slipshod in his ways.
74	N	When he does get over to work, he doesn't do a thing.
93	N	You have to explain every little step to him.
95	N	No cooperation at all.
113	N	Sometimes you can trust him and sometimes you can't.
115	N	Instead of doing it, he says, "I'll think about it."
120	N	He "goofs-off" all day long.

a reclassification on the basis of the present data would not represent a critical evaluation of the proposal, and it is too laborious to apply it to noncrucial data just because the latter are available.³

In the results of the several analyses reported herein, it is helpful to mention that all of them agree in differentiating more "poorest" descriptions from "best" ones at the lower levels of classification, where there are relatively many categories and each one classifies only a few individual patterns, rather than at the higher levels, where there are few categories and each classifies a large number of individual patterns under it.

II. A JOINT CONSIDERATION OF AGREEMENT AND FACTOR-ANALYTIC RESULTS

In this section types are interpreted in terms of results from both agreement and factor analyses. Then, the interassociations between factors and types are studied further in terms of items which are common and distinct for the types and factors.

Interpretation of the Types

Several of the primary advantages of agreement analysis are realized when we attempt to interpret the types. Each type is broad in the kinds of information furnished about the subjects who constitute the type, provided the test is wide in its coverage. This point is illustrated by listing the characteristics which were found in this study to belong to the first class (32 subjects) of "best" mechanics isolated. These characteristics are shown in Table 9, with the items which were answered "Yes" listed first, and those

answered "No" following.

The diversity of characteristics represented by these descriptions is illustrated by considering several of the items at the beginning of the above list. The first one tells about human relationships, the next one about quality of job performance, the next about reactions to questions, then about use of tools, personal traits, truthfulness, etc., in the order named, with each one giving information in a different area of behavior.

The diversity of information furnished about each type is further substantiated by a brief comparison of results from the agreement analyses with those from a square root, unrotated factor analysis of common data.

The 120 items which were analyzed here by two separate agreement analyses, using first 54 "best" mechanics and then 46 "poorest" ones, previously had been factor analyzed using 428 subjects (3), including those used in the agreement analyses. Using the results of this factor analysis, we examined the 43 items listed above as descriptive of the first class of "best" mechanics. These 43 items represent substantial loadings on 13 of the 23 factors reported in the factor analysis. Each of these 13 factors contains at least two items from the above 43 with loadings in its upper ten, and nine of the 13 factors contain four or more items (from the 43) with loadings in the top ten per factor. The first factor has all of its 10 highest loaded items represented in the 43 descriptions of the first class of "best" mechanics. The next two factors have none represented; the fourth has eight in the top ten. The others are as follows, with the factor number listed first, followed by the number of items from the first class with loadings in the top ten:⁴ 5-4, 6-5, 7-0, 8-5, 9-6, 10-0, 11-1, 12-5, 13-1, 14-7, 15-1, 16-5, 17-2, 18-2, 19-2, 20-0, 21-1, 22-1, 23-2. This comparison shows that the first class covers a wide breadth of information in terms of the number of factors involved.

Similar data for the last order of "best" mechanics (6 subjects) described in terms of 20 items, shows the following numbers of items

³ In addition, we have recently developed a method of multiple classification by agreement analysis whereby a subject can be classified in terms of all of the category patterns which are contained in his individual pattern.

⁴ Five of the factors had fewer than ten items listed because a lower limit for loadings of .100 was applied. These five factors had 9, 9, 6, 7, and 4 items respectively. Our comparison was restricted to the items listed.

TABLE 10

20 ITEMS DESCRIPTIVE OF THE FOURTH ORDER OF "BEST" MECHANICS ($N=6$)

Item No.	Response	Item
1	Y	He can get along with people pretty well.
14	Y	He has a nice appearance.
17	Y	He knows how to use tools.
36	Y	When he works, he really works.
68	Y	He makes sure he does a good job.
94	Y	He's just a "regular Joe."
97	Y	If you leave him to do a job, you can always be sure he will get the job done.
99	Y	He's just a normal guy.
107	Y	He is good at working on the plane.
119	Y	He seems to take pride in his work.
30	N	He doesn't have any sense of responsibility.
33	N	He thinks he has had a raw deal in the service.
35	N	He's just careless.
41	N	He achieves his aim in the wrong way.
53	N	You have to stand over him and show him how to do everything.
81	N	He shines in making decisions.
89	N	He occasionally drinks as he works.
95	N	No cooperation at all.
104	N	He isn't a credit to the Air Force.
120	N	He "goofs-off" all day long.

with relatively high loading on the factors, with the factors listed first, followed by the number of items: 1-3, 2-0, 3-1, 4-7, 5-1, 6-1, 7-1, 8-5, 9-2, 10-0, 11-0, 12-2, 13-1, 14-2, 15-1, 16-1, 17-1, 18-2,

19-2, 20-0, 21-0, 22-2, 23-1. Analogous data for the third family of "poorest" mechanics (8 subjects described in terms of 28 items), is as follows: 1-4, 2-4, 3-8, 4-0, 5-0, 6-6, 7-2, 8-0, 9-5, 10-2,

TABLE 11

28 ITEMS DESCRIPTIVE OF THE THIRD FAMILY OF "POOREST" MECHANICS ($N=8$)

Item No.	Response	Item
10	Y	He works in a sloppy way.
12	Y	You wouldn't feel safe unless you checked behind him.
23	Y	He is no "whiz" in his line.
27	Y	Most guys with that much experience know a lot more than he does.
38	Y	He isn't a very careful worker.
11	N	When he does a job, you know it will be done right.
16	N	He is one of those ideal mechanics.
19	N	Many times he works overtime even when he doesn't have to do so.
25	N	He knows what any modification of a Tech Order is about.
34	N	He is a good man on any job.
44	N	He deserves a promotion.
47	N	You don't have to worry about telling him what to do all the time.
48	N	He continues on without taking many breaks.
49	N	He will straighten a guy out and explain things to him.
50	N	He seems just about the best I've seen as far as getting his work done.
65	N	He will do an overtime job instead of leaving it to one of the others.
66	N	He always has better ideas than most men.
68	N	He makes sure he does a good job.
70	N	He's the energetic type who has to have something to do.
73	N	His work is nearly 100% correct.
79	N	He knows all the systems on an airplane.
87	N	If he were an instructor, he could put it over pretty thoroughly.
91	N	He knows his stuff.
97	N	If you leave him to do a job, you can always be sure he will get the job done.
100	N	He reads new Tech Orders at his first opportunity.
103	N	He keeps working as long as there's a man left standing.
105	N	He can show you how to do the job right.
106	N	His ambition will pay off.

11-0, 12-1, 13-0, 14-5, 15-6, 16-4, 17-1, 18-2, 19-0, 20-0, 21-8, 22-1, 23-0. These comparisons between factors and types show that the descriptions of types cover a wide variety of information in terms of the factors involved. An additional advantage is that information is given directly about specified subjects. It is not necessary first to analyze factors and then assess each individual in terms of the factors.

The gains realized in terms of an agreement analysis are not obtained without a sacrifice. The types cover such a breadth of information that it is difficult to do justice to the description of a type other than by listing the items which are descriptive of each type. We have already listed the items of a class (including the first two orders) to illustrate one type of "best" mechanics. Another type of "best" mechanic is illustrated by the last order of "best" mechanics isolated. The items of this order are listed in Table 10, with

those answered "Yes" listed first and those answered "No" following.

Normally, an order would have 16 individuals in a binary agreement analysis. The above order has only six because 48 of the 54 subjects in the sample were classified into the first three orders.

As an illustration of a type of "poorest" mechanics, the items descriptive of the third family (8 subjects) of poorest mechanics isolated are given in Table 11.

The items descriptive of the one additional order of "best" mechanics ($N = 16$), the one order of "poorest" mechanics ($N = 16$), and the three remaining families of "poorest" mechanics ($N = 8, 6$, and 8 , respectively) are listed in Tables 12-16.

The descriptions of the fourth family (Table

TABLE 12
31 ITEMS DESCRIPTIVE OF THE THIRD ORDER OF "BEST" MECHANICS ($N = 16$)

Item No.	Response	Item
2	Y	He wants to get in and do the job.
11	Y	When he does a job, you know it will be done right.
13	Y	He answers questions as best he can.
17	Y	He knows how to use tools.
31	Y	He always reports to work on time.
39	Y	You can always take his word for what he says he's done.
44	Y	He deserves a promotion.
47	Y	You don't have to worry about telling him what to do all the time.
57	Y	He sees decisions through to an end.
60	Y	He uses the right tools when he can get them.
68	Y	He makes sure he does a good job.
72	Y	He tries to find better ways of doing things.
84	Y	He stands up for what he believes in.
97	Y	If you leave him to do a job, you can always be sure he will get the job done.
101	Y	He usually remembers things which are explained to him.
105	Y	He can show you how to do the job right.
106	Y	His ambition will pay off.
107	Y	He is good at working on the plane.
109	Y	He likes being a mechanic.
118	Y	He gives good cooperation.
119	Y	He seems to take pride in his work.
10	N	He works in a sloppy way.
21	N	He is afraid to make a decision one way or another.
28	N	He just stands around until you're done before offering to help.
35	N	He's just careless.
53	N	You have to stand over him and show him how to do everything.
75	N	He doesn't seem to care what happens to him when he gets out of service.
95	N	No cooperation at all.
104	N	He isn't a credit to the Air Force.
113	N	Sometimes you can trust him and sometimes you can't.
120	N	He "goofs-off" all day long.

14) appear to be, in general, complimentary of the mechanics even though the mechanics were selected as "poorest" ones. This apparent inconsistency seems reasonable when we consider that the mechanics with whom we are concerned have been trained and selected for their jobs, and the terms "best" and "poorest" are relative and applied in this setting to small groups of mechanics. Each supervisor, who served as an informant in describing a mechanic, was requested to select either a "best" or a "poorest" mechanic and describe him. Each informant made his selection from those mechanics whom he had supervised in the last two years. Since the informants generally supervised not more than six or seven men in any one assignment, the group from which each informant made his selections was small. And each group, as already pointed out, contained only men who had been trained and se-

lected as mechanics. It is not surprising then to find that eight of the 46 men chosen as "poorest" mechanics are described rather complementarily. They probably are reasonably good mechanics. This conclusion from the experimental sample is consistent with the results from our cross-validation studies (Tables 5, 6, and 8) where at least 11 per cent of the "poorest" mechanics invariably classified as "best" ones in terms of our pattern analytic method.

However, there is a way in which the complimentary description of this family of "poorest" mechanics seems to differ from complimentary descriptions of "best" mechanics. The complimentary descriptions of "poorest" mechanics tend to avoid the most highly mechanic-centered descriptions. Instead, the compliments tend to be limited to the more general kinds of descriptions, such as: "He is a good man on any job"; "He

TABLE 13
38 ITEMS DESCRIPTIVE OF THE FIRST ORDER (2 FAMILIES) OF "POOREST" MECHANICS (N=16)

Item No.	Response	Item
9	Y	He accepts responsibility only when he must.
10	Y	He works in a sloppy way.
22	Y	His superiors should not give him a higher AFSC.
23	Y	He is no "whiz" in his line.
27	Y	Most guys with that much experience know a lot more than he does.
28	Y	He just stands around until you're done before offering to help.
30	Y	He doesn't have any sense of responsibility.
38	Y	He isn't a very careful worker.
51	Y	He does things without first finding out what will happen.
74	Y	When he does get over to work, he doesn't do a thing.
115	Y	Instead of doing it, he says, "I'll think about it."
11	N	When he does a job, you know it will be done right.
16	N	He is one of these ideal mechanics.
17	N	He knows how to use tools.
25	N	He knows what any modification of a Tech Order is about.
34	N	He is a good man on any job.
44	N	He deserves a promotion.
47	N	You don't have to worry about telling him what to do all the time.
48	N	He continues on without taking many breaks.
49	N	He will straighten a guy out and explain things to him.
50	N	He seems just about the best I've seen as far as getting his work done.
57	N	He sees his decisions through to an end.
64	N	Everybody likes him.
65	N	He will do an overtime job instead of leaving it to one of the others.
66	N	He always has better ideas than most men.
68	N	He makes sure he does a good job.
72	N	He tries to find better ways of doing things.
73	N	His work is nearly 100% correct.
79	N	He knows all the systems on an airplane.
83	N	There are times when he keeps working even though everyone else takes a break.
90	N	Everyone enjoys working with him.
91	N	He knows his stuff.
97	N	If you leave him to do a job, you can always be sure he will get the job done.
100	N	He reads new Tech Orders at his first opportunity.
105	N	He can show you how to do the job right.
107	N	He is good at working on the plane.
118	N	He gives good cooperation.
119	N	He seems to take pride in his work.

TABLE 14

22 ITEMS DESCRIPTIVE OF THE FOURTH FAMILY OF "POOREST" MECHANICS ($N=8$)

Item No.	Response	Item
6	Y	If he were a crew chief, he would work right along with his men.
20	Y	He doesn't just sit back with his hands on his hips.
34	Y	He is a good man on any job.
39	Y	You can take his word for what he says he's done.
47	Y	You don't have to worry about telling him what to do all the time.
50	Y	He seems just about the best I've seen as far as getting his work done.
60	Y	He uses the right tools when he can get them.
72	Y	He tries to find better ways of doing things.
84	Y	He stands up for what he believes in.
85	Y	What he doesn't know he can find out for you one way or another.
90	Y	Everyone enjoys working with him.
97	Y	If you leave him to do a job, you can always be sure he will get the job done.
99	Y	He's just a normal guy.
35	N	He's just careless.
42	N	He has unpleasant characteristics.
58	N	We have had a couple of "run-ins."
75	N	He doesn't seem to care what happens to him when he gets out of service.
82	N	He is just a kid.
93	N	You have to explain every little step to him.
108	N	He doesn't care to mix with other people.
112	N	He is pretty unhappy about being in maintenance.
115	N	Instead of doing it, he says, "I'll think about it."

stands up for what he believes in"; "He is just a normal guy." The "best" mechanics on the other hand are often described in terms of items which pertain more specifically to their particular jobs as aircraft and engine mechanics, such as: "He knows how to use tools"; "He is good at working on the plane"; "He seems to take pride in his work"; "He likes being a mechanic."

the first class of "best" mechanics are included in the 20 items descriptive of the last order of "best" mechanics. In other words, we have evidence here of two quite distinct types of "best" mechanics.

Only six of the 43 items descriptive of

the fourth family of "poorest" mechanics

TABLE 15

18 ITEMS DESCRIPTIVE OF THE FIFTH FAMILY OF "POOREST" MECHANICS ($N=8$)

Item No.	Response	Item
3	Y	He doesn't know his job.
22	Y	His superiors should not give him a higher AFSC.
23	Y	He is no "whiz" in his line.
27	Y	Most guys with that much experience know a lot more than he does.
31	Y	He always reports to work on time.
51	Y	He does things without first finding out what will happen.
80	Y	He has just an average working knowledge of the job.
84	Y	He stands up for what he believes in.
110	Y	It takes him quite a while to learn.
16	N	He is one of these ideal mechanics.
34	N	He is a good man on any job.
44	N	He deserves a promotion.
49	N	He will straighten a guy out and explain things to him.
50	N	He seems just about the best I've seen as far as getting his work done.
66	N	He always has better ideas than most men.
79	N	He knows all the systems on an airplane.
91	N	He knows his stuff.
100	N	He reads new Tech Orders at his first opportunity.

TABLE 16
7 ITEMS DESCRIPTIVE OF THE SIXTH FAMILY OF "POOREST" MECHANICS (N=8)

Item No.	Response	Item
9	Y	He accepts responsibility only when he must.
23	Y	He is no "whiz" in his line.
25	N	He knows what any modification of a Tech Order is about.
34	N	He is a good man on any job.
44	N	He deserves a promotion.
66	N	He always has better ideas than most men.
115	N	Instead of doing it, he says, "I'll think about it."

are included in the 43 items descriptive of the first class of "best" mechanics, and only two of the 22 items for the poorest family are included in the 20 items descriptive of the last order of "best" mechanics; *the numbers of items which are not common suggests that "best" and "poorest" mechanics differ in ways other than merely being opposite.*

Interassociations Between Factors and Types

In the above section on interpreting the types, comparisons were made with results from an unrotated, square-foot factor analysis of common data. In the present section, we wish to review the comparison for another purpose. We wish to call attention to interassociations between types and factors. The present evidence as to the exact nature of the associations is not maintained to be conclusive, and the data available are not the most crucial kind. The responses are descriptions by supervisors of mechanics, rather than known characteristics of mechanics. Consequently, whether the interrelationships derived from them reflect only the opinions of informants or are, in fact, characteristic of mechanics, is not known with certainty at this time. The factor analysis was unrotated, and by a square-root method rather than by a more exacting procedure. Likewise, the agreement analysis was by the binary

method rather than the more exacting and complete procedure (6, 7).

Of the three types of "best" mechanics reported, the large class of them (32 subjects and 43 items) is heavily involved in the first factor in that all 10 of the items with the highest loadings on this factor are contained in the 43 items descriptive of this type. Each of the next two factors (in order of amount of variance accounted for) have none of their ten highest-loaded items in the 43 items descriptive of the large class of 32 subjects. This class had a disproportionate amount of its variance accounted for in the first factor, and consequently other types were involved in the second and third factors. In the fourth factor, the large class reappears together with the other two types of "best" mechanics. Eight of the ten highest-loaded items on this factor are included in the 43 items descriptive of the large class, and five of these plus two others of the ten highest-loaded items are among the 20 descriptive of the last order of "best" mechanics. The relative association of these types with other factors can be seen in the figures reported earlier which show the number of high-loaded items for each factor from each type. Similar results were reported for the third family of "poorest" mechanics (8 subjects and 28 items). They reveal that this type is heavily involved in factors 3 and 21, with 8 high-loaded items from each. The two "best" types reported had only one item with a high loading in either of these two factors. The high degree of association between factor 21 and this "poorest" type is even more striking when we consider that this factor accounts for only 0.8 per cent of the variance in the data analyzed, and the agreement analysis of "poorest" mechanics was done on only 46 of the 428 subjects on which the factor analysis was performed.

In this comparison of types with factors we have illustrated our points with one class of 32 subjects and one order of six subjects from "best" mechanics, making a total of 38 "best" mechanics and leaving one order of 16 "best" mechanics

unaccounted for. Analogously, we have used only one family of "poorest" mechanics, leaving four families of eight mechanics and one of six mechanics unaccounted for. Data analogous to that already reported are summarized for all of the above-mentioned types in Table 17. This table reports the 23 factors which were isolated in the unrotated square root solution, listed in order of amount of variance accounted for. All of these factors except the 10th, 14th, 20th, 22nd, and 23rd had 10 items with loadings above 0.100. These specified factors had only 9, 9, 6, 7, and 4 items respectively with loadings above 0.100. Table 17 is concerned with the 10 highest-loaded items for each factor except those which had fewer than 10 above loading of 0.100. For these, it is concerned only with those above 0.100. The table also lists types of "best" and "poorest" mechanics derived from the agreement analysis. These are the particular ones used in the reliability study, reported earlier. For each type, the table shows the number of high-loaded items on each factor. The table reveals, as already illustrated, that *types* are associated with *factors*.

Table 17 contains an additional lesson. It shows a way in which factor and agreement analyses can be used jointly in interpreting behavior. For this purpose, four or more highly-

loaded items common to a factor and a type are taken to mean that a type is relatively high or low on a factor. For example, the four items listed under Factor 5 and opposite Type 1 mean that this "best" type is relatively high on this factor. Analogously, the four items under Factor 8 and opposite Type 4 mean that this "poorest" type is relatively low on this factor. The standings are all in the direction which would be expected from the names of the factors and the labels of "best" and "poorest" which are applied to the types, except for Type 6, the fourth family of "poorest" mechanics, which is complementarily described as we have already pointed out.

Applying the procedure just outlined, we see that Type 1 subjects are responsible; they are not lazy or weak in character. They do not fail to use knowledge effectively, are industrious, remember well, are socially acceptable, are practical workmen and do not lack craftsmanship. Type 4 subjects, on the other hand, are low on responsibility and willingness to work; they are lazy and fail to use knowledge effectively; they are not industrious and do not remember well; they are not practical workmen or craftsmen, and they lack both intellectual capacity and job knowledge. This "poorest" type is defined

TABLE 17
HOW ITEMS DESCRIPTIVE OF EACH TYPE ARE DISTRIBUTED AMONG THE
FACTORS IN TERMS OF THE HIGH LOADINGS^a

			Factor Titles and Numbers																						
Types	No. Items	No. Persons in Type																							
			1. Sense of Responsibility	2. Interest in Aircraft Maintenance	3. Willingness for Work	4. Laziness and Lack of Initiative	5. Weakness of Character	6. Failure to use Knowledge Effectively	7. Teaching Capacity	8. Industriousness	9. Memory	10. Self-Control	11. Inexperience	12. Social Acceptability	13. Lack of Morale	14. Practical Workmanship	15. Intellectual Capacity	16. Lack of Craftsmanship	17. Personal Pleasantness	18. Tendency to Mediocrity	19. Antisociability	20. Lack of Self-Control	21. Job Knowledge	22. Unlabeled	23. Unlabeled
"Best"																									
1—1st Class	43	32	10	0	0	5	4	5	0	5	6	0	1	5	1	7	1	5	2	2	2	0	1	2	2
2—3rd Order	31	16	2	1	1	0	2	4	0	0	4	1	0	2	3	2	2	1	1	2	2	0	0	2	1
3—4th Order	20	6	3	0	1	7	1	1	1	5	2	0	0	2	1	2	1	1	1	2	2	0	0	1	1
"Poorest"																									
4—1st Order	38	16	8	1	5	4	1	6	2	4	4	1	1	1	0	5	7	5	2	3	1	0	6	0	1
5—3rd Family	28	8	4	4	2	0	0	0	2	1	5	2	1	1	0	5	1	4	1	2	0	0	8	0	0
6—4th Family	22	8	3	0	1	0	0	0	0	1	1	0	1	3	2	2	1	0	1	2	4	1	1	1	1
7—5th Family	18	6	0	2	4	0	2	5	1	0	2	0	0	0	0	0	7	5	0	3	0	7	1	0	0
8—6th Family	7	8	0	0	1	0	0	0	0	0	2	0	0	0	0	3	0	1	2	1	0	0	0	0	0

^a For example, of the 43 items which describe the 1st class of "best" mechanics, 10 of them are the same items which have the highest ten loadings on factor 1; none of the 43 is among the first 10 in terms of loadings on factor 2; the same is true for factor 3; eight of the 43 are among the first 10 for factor 4, etc. All of the factors except the 10th, 14th, 20th, 22nd, and 23rd had 10 items with loadings above 0.100. These latter factors had only 9, 9, 6, 7 and 4 items respectively with loadings above 0.100. In the case of these factors, the comparisons were limited to the numbers of items just specified. In all other factors, the ten items with the 10 highest loadings were used.

as opposite to the above "best" type in all but two factors (weakness of character and social acceptability) of the eight used to define the "best" one. In addition, the "poorest" type is defined as deficient in three factors which were not used in defining the "best" type (willingness to work, intellectual capacity, job-knowledge).

In an analogous fashion, the other types can be described concisely from Table 17. In describing these types, it should be remembered that the raw data are descriptions given by supervisors; the patterns of descriptions may exist only in the minds of the supervisors or both in their minds and in the behavior of mechanics. Consequently, the patterns are not known at this time to be characteristic of the behavior of mechanics. They are here reported only as characteristic of descriptions of mechanics by supervisors. Additional research is required to determine whether they would also be obtained from objective assessment of behavior.

We have shown that some items are internally consistent in both factor and pattern analytic analyses. However, there were nine items which were internally consistent for some of the types by agreement analysis but did not appear with substantial factor loadings on any of the factors. These items are as follows:

- 6 If he were a crew chief, he would work right along with his men.
- 13 He answers questions as best he can.
- 15 He will climb into one of the airplanes and sleep.
- 34 He is a good man on any job.
- 55 If there is something he likes to do, he does it faster than anyone else would.
- 57 He sees decisions through to an end.
- 67 He is very glad to help someone else.
- 73 His work is nearly 100% correct.
- 75 He doesn't seem to care what happens to him when he gets out of service.

The one thing that these items appear to have in common is that they are *extreme statements*. It seems that some items can be highly dependable for describing some one or more categories of subjects even though they are not generally dependable for describing all subjects in terms of any one factor.

We searched for items which had relatively high loadings but were not included in any of the types on which the reliability study was done. (The search was made at an intermediate level of classification.) Relatively high loadings were defined as sufficiently high so that an item was included in interpretation of a factor (either in the first ten items in terms of loadings for a factor or a loading above .100 in the case of those factors which did not have 10 items with loadings above this value). There were 30 items with relatively high loadings (as just defined) which

did not appear in any one of the types on which the reliability study was done. In addition, there were 28 items with relatively high loadings which were not included in descriptions of "best" types and 25 other items with relatively high loadings which did not appear in descriptions of "poorest" types. These three categories of items plus the nine items already reported (which showed maximum relevancies for the types even though they had no relatively high loadings) make a total of 92 items on which there was considerable *difference* in the assessments of internal consistency by the two methods (agreement analysis and factor analysis).

Even though we grant that relative size of factor loadings on unrotated square-root factors, treated singly, is not a good estimate of internal consistency, the large discrepancies here reported show that there is a difference in the internal consistency of items as measured across all subjects (factor analysis) versus its assessment within selected categories of subjects (agreement analysis). In addition, these results show that many items differ in their validities for assessing factors and types. In short, these findings emphasize a new way in which reliability and validity are *relative concepts; they are relative to their assessment across all subjects versus their assessment within selected categories of subjects*.

The above results reveal the need for a more crucial comparison of results from factor and agreement analysis of common data, using the most refined technique of each method, together with data especially selected for the purpose of the comparison.

III. A CRITIQUE OF THE METHOD

It will be helpful now to react critically to the method just outlined and applied. The method is made computationally feasible by an arbitrary restriction; there is no reason (other than simplicity of classification) why the types at any one level should be composed of only two patterns from the next lower level. A more complicated version of agreement analysis avoids this arbitrary restriction (7). However, this latter method is laborious and it is felt that the present binary method is probably sufficiently exacting for our present state of knowledge.

Another problem of the method is that it sets no level of statistical significance which an agreement score must meet before a classification can be based on the score. The method circumscribes this problem by the following logic: It assumes types, and argues that if they exist for all patterns of response, then every pattern of responses must be classified and a solution is to give each pattern its best classification. This is a much less complicated solution than would be involved if the level of significance were required

to be computed for every classification of every pattern.

One could wonder whether or not "yes-no" answers are sufficiently dependable as a basis for classification. A reasonable reply to this kind of inquiry is to state that we do not have sufficient information of an appropriate kind on which to base an answer. Studies on the reliability of "yes-no" answers have been based on their dependability as assessed across all subjects of a sample. Agreement analysis recognizes the possibility of differential reliability of an answer; an answer may be highly reliable and dependable for some categories of subjects and not others. Agreement analysis enables us to determine whether or not "yes-no" answers have sufficient reliability of this kind to continue to justify the application of the method to them. Even if "yes-no" answers were found wanting, the method might still prove helpful if applied to categorized data, as obtained by dividing scores on several continuous variables into class intervals.

Agreement analysis does not allow us to assign every type a point in space in relation to every other type, and thus to give it a number which would reveal its location. Whether or not one wishes to do this depends on one's theory of the nature of types. If one wishes to assume that the same response in two different patterns has the same meaning, then it is reasonable to use the overlapping responses of patterns to assign types locations in space; this can be done in a method called linkage analysis (8). However, if one wishes to assume that the meaning of a response depends to a considerable extent on the pattern in which it occurs, then there would appear to be no basis for allocating the types to spatial locations on the basis of overlapping responses in their patterns.

Agreement analysis does not settle the differences in theoretical issues about types. It is, however, a method which should prove of assistance along with other methods in investigating some of the theories about types, and even in studying whether or not it seems worth while to assume types.

IV. SUMMARY

This paper has applied an abbreviated version of a method of analysis, called agreement analysis, for the isolation and classification of types into a hierarchical system. An unusual feature of the method is that the terminal table on which interpretations are based contains every bit of information represented in the raw data; information is abstracted and

generalized at successive levels, and the level most pertinent for a given purpose can be selected for further study. In addition, the method of analysis can handle all kinds of interrelations in data (linear, nonlinear, and disjunctive), and it also provides for giving various interpretations to any particular answer to an item, depending on the combination or pattern of other answers with which it occurs.

Illustrative applications of agreement analysis to descriptions of mechanics selected as "best" or "poorest" confirm the following hypotheses: The method yields both reasonably dependable and valid results. Descriptions of "best" mechanics can be differentiated from those of "poorest" ones on the basis of type membership. In addition, the results suggest the hypothesis that categories of people, such as "best" and "poorest" mechanics, can be most clearly differentiated in terms of many types, each type containing only a few highly similar subjects, rather than a few types containing many subjects. In terms of our data, there appear to be many types of "best" mechanics and many types of "poorest" mechanics—but no all-inclusive single type of "best" mechanic which differs from an all-inclusive type of "poorest" mechanic. Hypotheses of this kind can be further investigated by the method of agreement analysis which has been outlined here.

Results from our agreement analyses were compared with results from a factor analysis of common data. The comparison illustrated that the two methods classify data differently but not entirely independently; the factors and types which resulted are interassociated. Both methods yield results which appear to be meaningful and they can be used jointly in the interpretation of behavior.

However, some test items can have internally consistent responses for some categories of subjects in terms of an agreement analysis even though they are not sufficiently consistent across all subjects to yield any factor loadings of any consequence. These items appear to have value for selecting out rare subjects, such as a mechanic who is so expert that he is always 100 per cent correct; or, on the other hand, a poor mechanic who is so worthless that he would climb into an airplane and sleep.

Conversely, other items can be so in-

ternally consistent across all subjects that they have relatively high loadings and still not be retained as descriptive of any one of the predominant types at intermediate levels of classification. Consequently, the internal consistency of many items in terms of predominant types is quite different from their internal consistency in terms of factors. These issues on reliability of items under the two methods of analysis require further study by the most refined techniques of factor and agreement analyses with items especially chosen for the purpose.

REFERENCES

1. ETKIN, W. *College biology*. New York: Thomas J. Crowell, 1951.
2. MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.
3. MCQUITTY, L. L., WRIGLEY, C. F., & GAIER, E. L. An approach to isolating dimensions of job success. *J. appl. Psychol.*, 1954, 38, 227-232.
4. MCQUITTY, L. L. Pattern analysis—A statistical method for the study of types. In W. E. Chalmers, M. K. Chandler, L. L. McQuitty, R. Stagner, D. E. Wray, and M. Derber, *Labor management relation in Illini City*. Vol. II, Champaign, Illinois: Institute of Labor and Industrial Relations, Univer. of Illinois, 1954.
5. MCQUITTY, L. L. Pattern analysis illustrated in classifying patients and normals. *Educ. psychol. Measmt*, 1954, 14, 598-604.
6. MCQUITTY, L. L. A pattern analytic method derived from a theory of individual differences in psychological well-being. In S. B. Sells, Chairman, *Symposium on pattern analysis*. Randolph Field, Texas; Air University, USAF School of Aviation Medicine, 1955.
7. MCQUITTY, L. L. Agreement analysis: Classifying persons by predominant patterns of responses. *Brit. J. Psychol.*, 1956, 9, 5-16.
8. MCQUITTY, L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevances. *Educ. psychol. Measmt*, in press.
9. MEEHL, P. E. Configural scoring. *J. consult. Psychol.*, 1950, 14, 165-171.
10. ZUBIN, J. A technique for measuring like mindedness. *J. abnorm. soc. Psychol.*, 1938, 33, 508-516.

(Accepted for publication March 12, 1957)

Psychological Monographs: General and Applied

Contributions to the Study of the Problem-Solving Process¹ERWIN ROY JOHN²*Department of Physiology, University of California Medical Center, Los Angeles*

I. METHOD

PSYCHOLOGISTS have devoted a great deal of attention and effort to the investigation of human problem-solving, because of the theoretical value and practical utility of advances in this area of knowledge. Most of the studies in this area have focused on the success-failure aspect of performance, rather than on the internal structure of the performance. Behavior has been studied in terms of products rather than processes. The time

taken to achieve a product (i.e., to solve a problem), or the correctness of the product, or similar static indices have traditionally been the criteria by which performance was evaluated. A severe limitation on our level of understanding of such behavior has been imposed by the use of such methodologies; we are denied access thereby to information about the way in which factors within the performance itself contribute to the outcome, and the way in which these factors are related to other variables than problem-solving performance.

It is possible to achieve solution to a problem efficiently or inefficiently or even fortuitously. It is possible to fail brilliantly. In order to understand the dy-

¹The author wishes to acknowledge his indebtedness to a number of people who participated in various phases of this work. The Problem-Solving and Information Apparatus (PSI) was devised in collaboration with Dr. Horatio J. A. Rimoldi, now of the Department of Psychology, Loyola University, Chicago. Dr. Rimoldi also participated in the gathering of the data for the Ph.D. population.

The mode of analysis and scoring of PSI performances which is presented was developed with the help of Mr. Martin Balaban, of the Psychology Department of the University of Chicago, and Mrs. Margaret Labadie. The analysis of the data relating PSI performance to various personality measures was performed by Mr. Sidney Blatt, of the Psychology Department of the University of Chicago, who contributed in addition many ideas of value in the course of the research.

I wish to thank Dr. James G. Miller, of the Mental Health Research Institute of the University of Michigan, for enabling this research to be carried out, and for his enthusiastic participation in various phases of its development.

I am indebted to many of my colleagues for constructive criticism: Dr. Benjamin Bloom of the Examiner's Office of the University of Chicago, who gave me access to much useful data about entrance and course examination performance of many of my subjects, as well as

to his editorial ability. Dr. Joseph Kamiya, Dr. Martin Deutsch, Dr. Cynthia Deutsch, Dr. Samuel Sutton, Dr. Morris Stein, and Mr. Jo Banks have all aided my attempt to present this material with clarity.

Finally, I wish to acknowledge my debt to my wife, Dr. Vera John, for her assistance in formulating ways to examine the data, for criticism of the manuscript, and, most of all, for making available to me her knowledge of the area of human problem-solving research.

Some of the work included in this paper was done in connection with research supported by the Office of the United States Army Surgeon General under Contracts No. DA-49-007-MD-575 and DA-49-007-MD-684, Dr. James G. Miller, Principal Investigator. The opinions expressed in this publication are not necessarily those of the Office of the Surgeon General, Department of the Army.

²Performed this work while at the Committee on Behavioral Sciences, University of Chicago, and Mental Health Research Institute, University of Michigan.

namics of performance, to analyze the factors which relate to the achievement of a solution in a particular way, to investigate the possibility of relationships between variables external to the problem-solving performance and certain aspects of the process, it is desirable to make accessible for direct observation the problem-solving process itself. This requires the development of a technique which will permit the externalization of the dynamics of the process, such as the relationship between the product and the state of information of the subject, instead of continued reliance primarily on techniques which observe product and rely on introspection or subjective report to relate that product to its antecedent causes.

Recognition of this is implicit in some of the theoretical work in the problem-solving area by Duncker (2) and others who have attempted to analyze problem-solving as a process in order to understand the relationships between successful performance and factors within the behavior itself. Theoretical and conceptual schemes have been developed about typical behavior in various situations, and facilitation or hindrance of success. These attempts have been constrained by reliance on recorded verbalizations or qualitative descriptions by an observer of the behavior. In recent years, there have been a number of promising attempts to develop an experimental methodology which would permit more operational definition of the elements of behavior and more precise quantification of these elements. To the extent that such attempts are successful they will enable development of theoretical schemes which can be subjected to experimental verification.

Recognition of the necessity and the desirability of analyzing process rather than product is explicit in the excellent monograph on problem-solving by Bloom (1). The reader is recommended to read this work, if he is not already familiar with it, since it contains not only cogent arguments for preoccupation with devising process measures but a large number of experimental findings which are confirmed in more quantitative terms by the present work.

Some attention has been given to various versions of the game of Twenty Questions, by Harlow (5) and a number of other workers, as a source of process information. These approaches

have arisen from concern with military training devices or trouble-shooting analyses, and have tended to deal with procedures for the repair of faulty equipment of one sort or another, where checking procedures are assumed to be binary operations which eliminate each time one-half of the remaining alternatives. While such inquiries are of interest, their utility tends to be limited by the structure of problems which permit such binary operations. Of perhaps more general interest are techniques developed for very similar purposes but without this restriction in structure, such as the Multiple-Alternative Symbolic Trouble Shooting Test (MASTS) of Grings et al. (4), which is yielding some information about previously inaccessible aspects of the trouble-shooting process, in electronic problems. Again, the highly specialized nature of the task limits the utility of results so far obtained for understanding of more generalized problem-solving processes.

Of greater generality, and therefore perhaps of greater potential utility in the investigation of the problem-solving process, is the Tab Item technique, recently developed by Glaser et al. (3). While this task was developed for the analysis of electronic trouble-shooting proficiency, like those above, it possesses a flexibility which permits its application to many other areas of inquiry. The technique consists of presenting a subject with a problem and a set of labeled items of information potentially of utility in the solution of the problem. These items are selected by the subject in any order which seems pertinent to him, until he has gathered those items which he requires for the solution of the problem. The sequence of selection of these items by the subject is used as the index of the way in which already acquired information is used. The authors of the Tab Item indicate their recognition of the flexibility of the technique in areas of more general concern than electronic trouble shooting when they say "Proficiency in medical and psychological diagnosis might be tested by adapting this technique to measure the ability of a clinician to perform the procedures necessary for correct diagnosis. . . . In general, the Tab Item is an applicable technique for the measurement of behavior which involves the serial performance of a set of procedures where the performance of one procedure yields information which supplies a cue for the selection of the next and subsequent procedures." Rephrased, this technique might be of utility in the investigation of problem-solving where the information derived from examination of one aspect of the problem can be used to direct subsequent inquiry to specific other aspects of the problem until those aspects necessary for a stipulation of the solution to the problem are understood.

Working independently, Rimoldi (7) has developed and applied a technique very similar to the Tab Item for the investigation of the process of solution of problems of medical diagnosis and chemical analysis. As in the Tab Item, this technique consists of presenting a problem to a subject and then making available a body of information, containing items which are more or less relevant to the solution of the problem. These items are labeled so that information concerning specific aspects of the problem can be elicited as the subject chooses. The order and nature of the items selected by the subject are recorded.

While the tests referred to above represent a large step forward in problem-solving methodology, they nonetheless appear to the present writer to suffer from a number of drawbacks. As used to date, they are restricted to problem-solving which requires a high degree of specialized knowledge, with the consequence that the process observed is dependent on the education of the testee to an undesirable extent. In addition, it is exceedingly difficult to stipulate the information content of the items available to the subject, or to equate the information content of these items, or to evaluate the extent to which each item contributes to an understanding of the problem as a whole. Perhaps most important, *the format of these problems is not realistic. They tend to structure the path of the subject to an undesirable extent by supplying a limited number of obvious alternatives, give away information by listing the possible alternatives and crucial cues and checks which can be performed, and remove the additional dimensionality which results from having the subject actually deal with real events, rather than their abstract paper representations, in a situation where repetition is possible, and where the set of possible actions is not artificially delimited or ordered in time.*

For greater utility, the ideal format for observation of the problem-solving process should start the subject with a standard minimum of information about a problem and then require him to structure his own presolution behavior with a minimum of externally imposed constraint. Such a format should be based on a task which is maximally free of special skills, special knowledge, or experiences peculiar to a given culture. Since it is important that repeated administration of comparable forms be possible, so that changes in process after

various kinds of intervention can be measured, it is also desirable that the effect of familiarity with the generic task be minimal. It is vital to be able to quantify the difficulty of the task, and to be able to vary the difficulty over a large range, as well as to know the information contained in any element of the behavior. Finally, it is desirable to devise a format which presents the subject with the necessity to interact with real events rather than abstract.

A technique which permitted the direct observation of the problem-solving process and met the above stipulations would offer a number of advantages. It would permit analysis of the problem-solving process itself to see whether it was qualitatively homogeneous, or whether it is composed of qualitatively distinct phases which could be defined and described. Theoretical formulations, such as those of Duncker, for instance, could be evaluated against such information and modified if needed. The factors determining the relationship between such phases, and affecting the transition from phase to phase, might emerge from such observation. The relationship between education, experience, cultural background, personality variables, and the internal factors which interact in the problem-solving process might be clarified. The extent to which particular aspects of the process are characteristic of an individual might be determined. The reorganization in process which might result as a consequence of change in the severity of the demands of the task on the individual could be analyzed. The extent to which various techniques or experiences are of value or of utility in increasing the effectiveness of an individual's problem-solving behavior might be evaluated, and knowledge gained about how to change specific aspects of the behavior. Relationships between such aspects of behavior and skills or abilities involved in real-life activities might become evident, enabling better direction of the acquisition of these skills. Diagnostic examination of the problem-solving behavior of an individual might permit application of remedial techniques to the areas of process which most constrain that individual's performance.

The ultimate realization of these possibilities would require extensive research, once suitable instruments for the observation of process were devised. The achievement of the ideal format described earlier represents a difficult task in itself.

This paper describes a technique, briefly reported previously (6), which it is felt constitutes a step in the correct direction. While numerous criticisms can be directed against this technique in that it does not fully meet some of the desiderata, yet it provides much information previously inaccessible and should be of use in the development of still more appropriate methods, and in the acquisition of basic knowledge about problem-solving processes.

In considering the results obtained to date using this technique, reported subsequently in this paper, we must emphasize that we are here dealing with problem-solving performance of a particular kind. *In these problems, a finite and enumerable set of discrete relationships must first be elucidated and then manipulated so as to achieve a single stipulated result.* There clearly are problems which do not fall in the category represented by our technique; problems such as those involved in certain kinds of research in the sciences would appear to be rather similar to our generic task, while problems such as those involved in the making of artistic products, where there is no right answer, or the elucidation of less discrete relationships using probabilistic evaluations, would appear to be quite dissimilar, although perhaps analogous methods can be devised for their investigation. One should not fall into the error of assuming that performance on this task represents evidence of some sort of basic ability which is involved in all sorts of productive activity. This word of caution is made necessary by the many examples of misapplication of "tests" of reasoning ability in selection procedures currently.

Further, there is an unquestionable element of stress and competitiveness inherent in the mode of presentation of the

task, as will be described in the following pages. Evidence that subjects are susceptible to this stress and consequently display a performance which by no means represents their optimal behavior has been gathered in a preliminary experiment, and imposes a further constraint on conclusions about the relationship between "real" ability and the sample elicited in our task.

While we will present evidence that behavior in the situation which we have constructed can be interpreted as an approximation of the subject's habitual approach to problems which fall in this category, and is meaningful as such, yet it is apparent that by appropriate instruction or experience over a period of time, or by diminution of the threatening aspects of the situation, this habitual approach can be modified.³ The extent to which such modification would be generalizable to analogous real-life tasks is a problem of great practical importance.

However, in our desire to qualify the interpretation of results gathered using this method, let us not obscure our belief that we have here defined a real and important class of problem, which would appear to be relevant to a number of

³ Three individuals were tested in the PSI situation separately on one problem. They were then brought together into a group and asked to solve a second problem. The procedure was for each individual to hand to the experimenter the next step which he would make if he were working alone, and then for the group to discuss its next move until unanimous agreement was achieved. The situation compelled the subjects to formulate verbal arguments for their proposals which would persuade the other members of the group. In this situation radically different behavior was exhibited by the group members from their previous performances, which did not compel the use of language with the consequent imposition of syntactical structure on their "reasoning." This experiment was carried out with the collaboration of Mr. Richard Mann of the Mental Health Research Institute of the University of Michigan.

real-life situations, and that this method gives access to information about problem-solving behavior previously unavailable.

Description of the Problem-Solving and Information Apparatus (PSI)

Problems can be conceived of as consisting of two aspects: elements and relationships. The logic of the relationships between elements can be stated in terms of symbolic logic, devoid of factual content. Thus it is possible to construct an abstract logical problem with the same formal relational structure as a problem which involves factual content. Any logical relations, or propositions, which can be stated in symbolic logic can be stated as an equivalent electromechanical circuit. It is possible, if one confines oneself to problems of a structural complexity commensurate with the number of circuit relationships available, to construct an apparatus on which a large number of sets of logical propositions can be stated. The difficulty (complexity) and structure of these sets of propositions can be quantified as desired, and easily varied by the experimenter. A given set of propositions constitutes a "problem." The achievement of a particular output from a network of electromechanical circuits, which are related in accordance with such a set of logical propositions, requires that the constraints imposed by the logical relationships as defined by that set of propositions be satisfied. The achievement of such an output constitutes a solution to the problem.

In accordance with the foregoing considerations, an apparatus has been constructed (6) and a technique devised which permits observation of aspects of the problem-solving process which fall into three categories:

1. Factors which relate to the amount of effort required and the characteristic habits of work displayed in the process of achieving the desired output from the apparatus.

2. Factors which relate to the acquisition and handling of information in the process.

3. Factors relating to the organization, manipulation, and synthesis of acquired information in order to achieve the desired output.

This apparatus will be designated as the Problem-Solving and Information Apparatus (PSI) in what follows.

Using the PSI, numerous aspects of an individual's problem-solving performance can be defined and measured quantitatively in each of these three categories. The definitions of these variables will be presented in the next section.

Information theory permits the precise quantification of the total information content, or "difficulty," of any set of propositions, or "problem." One can construct a number of problems of the same or graded difficulty. Such problems can be administered serially to ascertain the effect of increasing difficulty, or the extent to which learning occurs in this situation. The technique permits sequential analysis, making possible a treatment of problem-solving as a branching stochastic process with nonindependent transition probabilities. In other words, one can determine the extent to which the sum of all the information obtained by a subject in the first $n - 1$ steps of a given performance influences the n th step. The utility of this will be discussed further in the section on variables.

The PSI itself consists of a set of electromechanical elements arranged, with other components, into a Boolean computer. Desired logical relationships between the elements can rapidly be established by means of a plugboard in the rear of the apparatus. A given set of relationships defines a problem. Problems are constructed so as to have a unique

solution. A solution is defined as the production of a particular output from the network of related elements by using three particular elements, called "input elements," in some combination or temporal sequence. In other words, the apparatus can be conceived of as composed of three divisions: input elements, an output, and a network relating the input elements to the output. The task of the subject, then, is to learn how to achieve a stipulated output from the network of related elements, by using only the input elements. This requires that he analyze the logical relationships between the elements and then utilize these relationships to produce the desired output. In all of the problems constructed to date, the required output and the set of input elements have been the same. Problems differ only in the structure of the network of relationships between the input and output. Front and rear views of the apparatus are presented in Fig. 1.

On the face of the PSI is a display panel. On this panel is presented a circular array of nine pushbuttons paired with a circular array of nine lights. There are nine basic electromechanical elements in the network, and each light indicates the state of one of the elements. The corresponding pushbutton permits the activation of that element. Inside the circle of lights is a disc on

which is presented an array of arrows connecting various elements. An arrow between two elements indicates the existence of a relationship between those two elements, with the direction of the relationship indicated by the head of the arrow. All arrows on the disc stand for relationships in the logical network, and all relationships which exist in the logical network are indicated by arrows on the disc. In the center of the disc is a white light, which is the desired output from the network. Clearly, to each particular problem, defined by a set of relationships, there corresponds a particular problem disc.

Although the existence of all relationship is indicated on the disc, the specific nature of the relationship is not. An arrow from "A" to "B," for instance, might mean (a) that A was sufficient to cause B, or (b) that A was necessary but not sufficient to cause B, or (c) that A was sufficient to prevent B. Sometimes the nature of a relationship can be uniquely inferred simply by inspection of the disc, knowing that all relationships indicated exist and that all relationships which exist are indicated, together with the given fact that all problems are soluble. Sometimes inspection alone does not permit the drawing of a valid inference without ambiguity. In such a case, the subject must devise a procedure to elicit the necessary information from the network. The subject may ascertain the nature of any relationship he desires by activating the pertinent elements by means of their associated pushbuttons and observing the series of consequences of this activation as displayed by the lights on the panel. In other words the subject designs small logical experiments, from the interpretation of which he may infer the nature of the relationships among the elements. *He may use all the elements as many times and in as many combinations as he desires, in order*

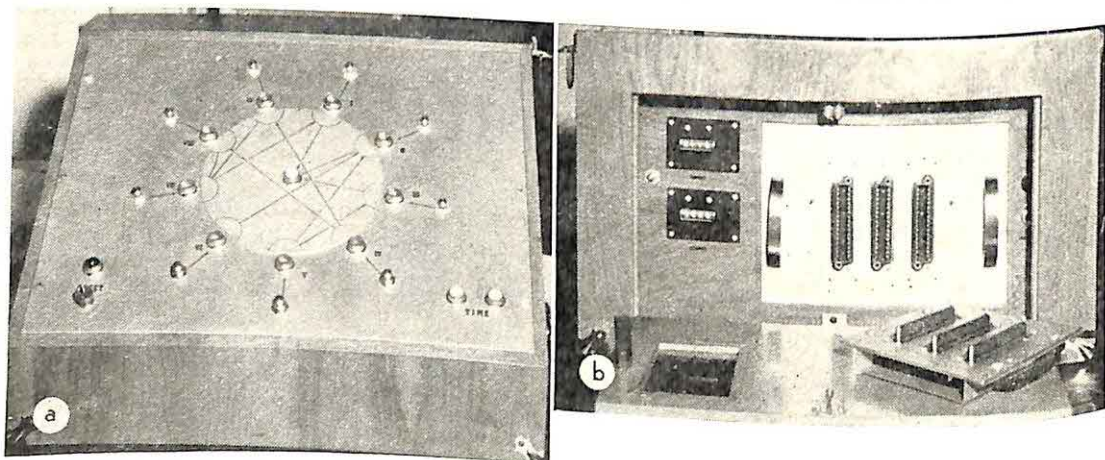


FIG. 1: a. Front of PSI apparatus, with example disc in place. b. Rear of PSI apparatus, showing plugboard and problem ready to be inserted.

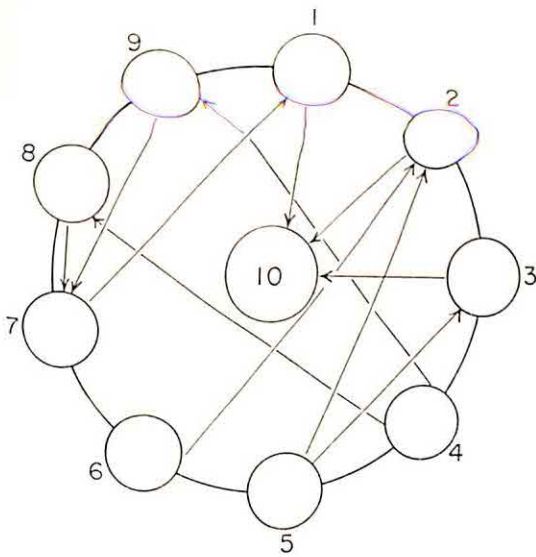


FIG. 2: Example problem disc.

to obtain the information which is necessary to permit the achievement of the required output using only the permitted input elements. Observation of the order, nature, and time of these experiments is the raw data for the sequential analysis referred to above and described in the next section.

Perhaps an example may be useful to clarify the task which the subject faces and to illustrate the way in which information can be elicited from the apparatus.⁴ Figure 2 shows the problem disc which has been used to date to accomplish familiarization of the subject with the PSI.

The task of the subject is to learn how to light 10 by some combination or sequence of manipulations which involve using only 4, 5, and 6. In order to learn what this combination or sequence should be, the subject may ask questions using any of the 9 elements in any order or combination. He may also, if he wishes, not ask questions which use elements other than 4, 5, or 6, and attempt to achieve 10 simply by permutations of these three elements. In the procedure which was used in the work reported later in this paper, subjects were asked to achieve solution in minimum time and using minimum questions of the machine, in which case permutation is inefficient, although permitted, behavior.

Let us work the example to illustrate one way in which the subject might proceed. There are three arrows which lead to light 10, the

required "output." They come from 1, 2, and 3. These arrows might mean that 1 alone, 2 alone, 3 alone, 1 & 2 but not 3, 1 & 3 but not 2, 2 & 3 but not 1, or 1 & 2 & 3 together is the necessary and sufficient condition (NASC) for 10 to occur, in accordance with the permitted relationships as enumerated on page 6, clearly one of the foregoing must be true. In order to ascertain which of these possibilities applies in this case, we must ask questions of the PSI. Pressing the pushbutton 1 on the panel will cause the light which indicates the state of 1 to be lit on the display panel. If 1 were the NASC for 10, after three seconds 1 would go out and light 10 would be lit on the panel. If 1 is not the NASC for 10, after three seconds light 1 will go out and no light will be lit on the panel. In our example, the NASC for light 10 to be lit is 1 & 2 but not 3. This means that 1 & 2, if lit together, will cause 10 to light if 3 was not also lit. By the appropriate manipulations of the PSI, the NASC for 10 are thus determined. From this point on, it is possible to solve the example without the necessity of asking any further questions. Let us see how this can be accomplished, and at the same time let us illustrate how the necessary information for the solution can be achieved by asking questions of the PSI, if one does not draw the inferences already possible.

Let us assume that we have asked the necessary questions to learn that 1 & 2 but not 3 is the NASC for 10. We now must learn how to light 1 & 2 without 3 by some combination or sequence which involves only 4, 5, and 6, to meet the criterion for a valid solution. Let us first consider how to achieve the ability to light 1 in this way. Only one arrow comes to 1, and it comes from 7. If 1 can be lit by any element, that element must be 7. Since this example, like that permitted problems, is stated to have a valid solution, 7 must cause 1. If we do not draw this inference, we can ask what the arrow from 7 to 1 means by pushing button 7. Light 7 will be lit, and three seconds later will go out as light 1 is lit. Three seconds later light 1 will go out, and 10 will not be lit, because 1 is necessary but not sufficient for 10. The problem of how to light 1 can now be restated as the problem of how to light 7, since we have learned that 7 is sufficient to cause 1. Two arrows come to 7, one from 8 and one from 9. This might mean that 8 alone, 9 alone, or 8 & 9 together, is the NASC for 7. We can, if we wish, ask the PSI which of these is true by pushing the various combinations and observing the results. However, if we look one step further, we see that only one arrow comes to 8 and only one arrow comes to 9. Both of these arrows come from 4. If the arrow were to stand for the relationship prevents, it would be meaningless because no other element is permitted to

⁴ PSI is commercially available, with or without automatic recorder, from Standard Electronics Company, 6728 S. Halsted Avenue, Chicago, Illinois.

cause 8 or 9 within the rules, since no other arrows are indicated. We have stipulated that all arrows on the disc must stand for real relationships. Thus we can rule out the possibility that 4 prevents 8 or 9. Similarly, 4 can not be necessary but not sufficient for either 8 or 9 because there is no other arrow to these elements. Thus, 4 must cause *both* 8 and 9. Now, if it is impossible to achieve 8 without 9, and vice versa, we can rule out the possibility that either one can prevent the other from causing 7, because if that were the case this problem would have no solution. Thus we conclude that 8 & 9 is the NASC for 7, which is the NASC for 1, and we can achieve 8 & 9 from 4.

We have now learned how to achieve one part of the NASC in a fashion which is permitted in the requirement for a solution, which the reader will recall prohibits the use of any other elements than 4, 5, and 6. We must now learn how to achieve 2 without 3, in a similar fashion. Two arrows come to 2, one from 5 and one from 6. This might mean that 6 *alone*, 5 *alone*, or 5 & 6 *together* is the NASC for 2. We can ascertain which of these relationships actually holds by asking the necessary questions of the PSI, or we can take an additional factor into account. We know that 3 will prevent 1 & 2 from lighting 10, since the NASC for 10 is 1 & 2 *but not* 3. Only one arrow comes to 3, and it comes from 5. If this arrow meant *prevents*, the relationship would be meaningless, since no other element can cause 3. Similarly, it can not mean that 5 and some other element are together necessary for 3 to occur, because there is no other arrow to 3. Thus, 5 must cause 3. If this is true, then if 5 & 6 *together* is the NASC for 2, 2 will always be accompanied by 3, and the problem cannot be solved. Since we have stipulated that the problem is soluble, 6 must be the NASC for 2. We can not infer validly whether or not 5 also is sufficient for 2, or whether 5 prevents 6 from causing 2, on the basis of the information we possess. This information is irrelevant, since we can solve the problem without the necessity of resolving this ambiguity. Of course, the ambiguity can be resolved simply by pushing 5 alone and observing the consequence, which is that 2 is lit by 5 as well as by 6.

We now know the meaning of every relationship in the network. We know how to achieve 1 using only the elements permitted in a solution, and similarly we know how to achieve 2 without 3. To achieve 1 we must use 4, which will cause 8 and 9, which will cause 7, which will cause 1. To achieve 2, we must use 6, not 5, because we must prevent 3 from accompanying 2. However, we must achieve 1 and 2 at the same time, since 1 and 2 together is the NASC for 10. A moment of thought will make clear to the reader that if 4 and 6 are pushed together, after three seconds 8 and 9 and 2 will be lit, then 2 will go out and

8 and 9 will cause 7, and when 7 causes 1, 2 will already have gone out. In other words, we have a temporal aspect to this problem. We must arrange our manipulation of 4, 5, and 6 into a sequence which will give us the various parts of the NASC *at the same time*. In order to do this, we must use 4 to cause 8 and 9 to occur, wait until 8 and 9 have caused 7 to occur, use 6 *at that time* to cause 2 at the same time that 7 causes 1, and now we have established the NASC in a way which meets the restrictions imposed by the definition of a valid solution.

The example contains all the kinds of relationships of which problems on the PSI are composed. In the example, it is not necessary to ask further questions once the necessary and sufficient condition (NASC) for 10 is known. Not all problems, of course, have this characteristic. An enormous number of different problems can be constructed from the various meaningful combinations of these relationships which the design of the PSI permits. The information content of these problems can be rigorously evaluated, and the content of each possible question which a subject might ask of the PSI can also be evaluated, in a fashion which will be discussed in detail in the following pages.

It should now be apparent that the subject, as a result of his manipulations of the pushbuttons in the course of investigating relationships, is presented with ordered sequences of events which are consequences of his activation of elements. His task is to extract, from observation of the sequences which he designs, that information which he needs in order to be able to ascertain the characteristics of a valid solution to the problem. Such a solution may require an exceedingly complex ordering and combination of the input elements.

Note: The three-second delay which ensues before the consequences of any antecedent are displayed has been imposed to give the subject the opportunity to see the full set of consequences of any change of state which he initiates,

as an ordered series of events, and as well to afford sufficient time for him to execute any manipulations which he may desire to add at any point in the series. In preliminary studies, a seven-second delay was provided and subjects complained about the excessive delay. A three-second delay seems empirically to provide the opportunity to order events adequately without causing the subject to be annoyed by having to wait, and is considered a satisfactory compromise. Because of the large variation in rate of manipulation of the PSI, described under *Rate* in the section on variables which follows, it is felt that the three-second delay is not the cause of the various rate phenomena which have been found. It is, of course, possible to check this point by providing a manually controlled delay operated by the subject at will, but this has not yet been done.

Some readers may be familiar with the models of neural nets which have been extensively used by neurophysiologists and cyberneticists. It may be of assistance to conceptualize the PSI as an electro-mechanical neural-net model, in which direct excitation, summation, and inhibition can be combined in any desired fashion. Or, perhaps, the reader may appreciate further clarification from Fig. 3, which is a verbal equivalent of a problem on the PSI, as it could be represented once the subject has elicited all the information necessary to stipulate each relationship. Note that once the subject has obtained knowledge about all the relationships he is still left with the nontrivial task of utilizing this knowledge to produce a solution of the problem as required by the rules.

Familiarization with the PSI is achieved by stating the rules, describing the nature of the possible logical relationships which the subject might expect to encounter later, and then illustrating all of these by presenting the comprehensive detailed example which we have just completed. The subject is provided with paper and pencil, and encouraged to take notes. Young children are able to work effectively on the PSI, indicating that the situation itself offers little diffi-

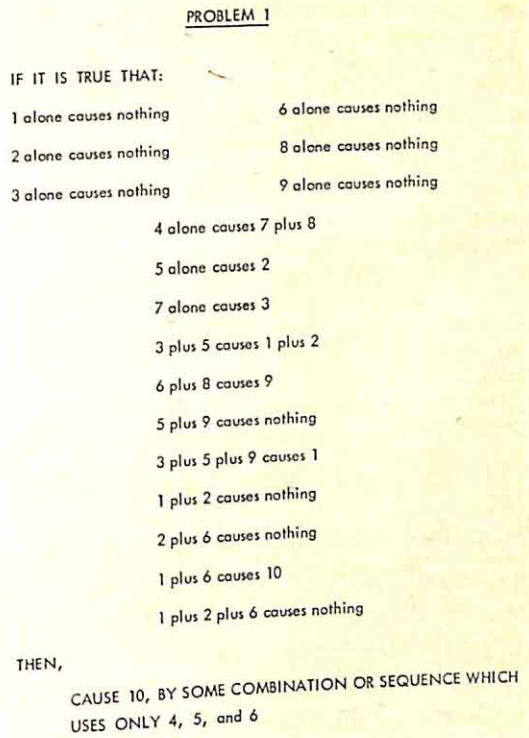


FIG. 3: Verbal equivalent of first problem. Note: This represents the problem once all relationships have been analyzed.

culty of comprehension when illustrated concretely instead of in the manner used above.

II. PROCEDURE AND SCORING

A. General Considerations

A large amount of data is made available from a single performance on this apparatus. There are many possible ways of organizing and examining these data. As will be apparent from the subsequent list of variables, it is not difficult to define a variable which is a new way of looking at some aspect of the data. The possibilities for such new variables have by no means already been exhausted. Particularly, the definition of variables which give access to the dynamic aspects of our data without emphasizing only the static aspects has been difficult. Our habitual approach has been in terms of

products rather than process, and, presented with precise and detailed data about process, one tends still to describe the product rather than the process. It is hoped that, as work in this direction proceeds, a set of more adequate process indices will be developed.

Many of the variables defined are reasonably independent, many covary appreciably with others. In view of the differences in performance of groups of diverse training and background, which will be illustrated in the data to be presented, it would appear unprofitable to investigate the intercorrelations between these various variables at this time. The evidence suggests strongly that one should use a homogeneous population in order to determine intercorrelations between the variables, in view of the probability that the internal organization of the process, and therefore the correlations to be achieved, will vary as the structure of the group varies. Changes in situation and instruction may also change the internal organization. This pilot study did not attempt to obtain such information. Perhaps ultimately factor-analytic methods should be used to select a set of variables which span the space of the performance. However, the intuitive meanings and significance of many of these variables have been so useful conceptually that we have made no effort to prune our set of variables to a minimum. Such an attempt would appear premature until an appreciable period of examination of the many different methods of organizing this somewhat unfamiliar kind of data has elapsed.

The variables so far defined and examined can be grouped into three different major aspects of performance:

1. Those variables which describe the individual's work habits and relate to the effort ex-

pended in the solution of the problem are referred to as *work variables*.

2. Those variables which describe the way in which the individual acquires and handles information are referred to as *information variables*.

3. Those variables which describe the orientation of the individual to the problem, and which relate the effect of acquired information to this approach, are referred to as *approach variables*.

B. Problems

Subjects were presented with two problems on the PSI. Both problems were administered to each subject separately in a single session, which usually lasted about one and one-half hours, including time for familiarization and demonstration of the example.

The two problems were constructed so as to be almost identical, with one major difference distinguishing them. Each contained 17 items of information in the total information pool; that is, 17 simple propositions would define the network of relationships exhaustively. Each problem contains 4 relationships which are direct implication, 2 relationships which are conjunction, 1 relationship which is compound conjunction and disjunction (x will occur if and only if y and z but not w has occurred), and 1 relationship which is disjunction. The difference between the two problems is that the first problem requires that *three* coincidences be achieved in the course of the solution, while the second problem requires that *four* coincidences be achieved in the course of solution. One of the four coincidences used in the second problem is the same relationship which is required to produce a prior coincidence, but this relationship is used for a different purpose in the two instances. Thus the difference between the two problems might be summarized by stating that they are identical except that a relationship used only once in the

first problem is used in two different ways in the second problem, as a consequence of using different elements in the same relations.

Attention is directed to the apparent small difference between these two problems, since presumably the differences in performance observed are a consequence of these minimal differences in structure. Yet *this minimal difference in structure, achieved without any increase in the number or nature of the items in the pool, brings about a major change in the behavior elicited from our subjects.* The data presented below give us an insight into the extent to which small changes in structure may change aspects of response which superficially one would expect to remain relatively unaffected. Clearly, the *structure* of relationships, as well as the sum of the information contained in these relationships, is a major determinant of the ease with which *meaning*, or coherent organization of the set into a whole which is unified, can be achieved. (One might suppose, since alternative structures may lend themselves to the same content, that clarification of this point might generate understanding of the factors which constrain comprehension under various conditions of the presentation of information.)

C. Instructions

In the experiments which yielded the data to be here reported, subjects were asked to solve the problem in a *minimum* of time and with a *minimum* number of inquiries of the PSI. The evidence strongly suggests that other instruction, for instance to minimize time alone or *inquiries alone, or other situations, which increased the probability that the subject would use particular mediational processes such as language, might lead to a different approach to the situation.*

This particular choice of instruction was made in an attempt to place the subject under some pressure to perform effectively while at the same time minimizing the extent to which the situation was structured. It was hoped that this would enable us to observe the way in which our subjects themselves tended to structure cognitive tasks of a highly abstract nature.

It should be clear that the PSI permits the construction of other situations, and one can then compare the behavior elicited in these situations to that obtained under the conditions cited above. Specific content can be put back into the problem, instead of using an abstract presentation. Problems can be presented with all relationships indicated and their nature specified. Problems can be presented with none of the relationships indicated. Networks can be constructed which are not constant, but which change as a function of certain aspects of an individual's performance. Other instructions can be used. Betting behavior, or behavior under stress, or under various kinds of motivation can be observed. Group situations can be studied. Much can be learned about the dynamics of problem-solving behavior, and other kinds of behavior as well, by such variations. None of these interesting alternative approaches has been studied to date. In view of the many possible variations in the situation in which the PSI is administered and in the amount of content introduced into the particular problems selected, the information which has been gathered to date must be regarded as surveying only a small part of the domain which can be investigated using this basic technique. We will develop certain ideas and approaches to the analysis of this kind of sequentially observed behavior which may be of utility in the direction of subsequent studies. However, the results obtained with regard to differences between the performance of various kinds of groups and with regard to characteristics of the problem-solving process must be regarded as gross outlines which must yet be filled in, rather than as definitive descriptions. This is particularly true in view of the small size of our various groups.

D. Subjects

The subjects in these experiments were University of Chicago students and staff members who responded to an advertisement offering to pay persons who

would participate in a psychological experiment. These students varied appreciably in their length of attendance at the University and came from various academic disciplines. On the basis of the criteria stated below, they were divided into groups as indicated.

Group 1: 21 Ph.D. candidates or recent Ph.D. recipients. (Composed of Group 3 and Group 4, below.)

Group 2: 16 students in their first quarter of residence in the College of the University of Chicago.

Group 3: 10 candidates for or recent recipients of the Ph.D. degree in the natural sciences.

Group 4: 11 candidates for or recent recipients of the Ph.D. degree in other fields (8 social sciences, 3 humanities).

Group 5: 11 first-, second-, and third-year students in the College of the University of Chicago who indicated in an essay on career plans their intention of studying for advanced degrees in the natural sciences.

Group 6: 11 first-, second-, and third-year students in the College of the University of Chicago who indicated in an essay on career plans their intention of studying for advanced degrees in areas other than the natural sciences.

The mean performance of each of these groups will be presented for each PSI variable on both of the problems described in Section B. Our purpose in the remainder of this chapter is to present the definitions of the variables which constitute our present mode of analysis of PSI data, and to gain some insight into the extent to which the desiderata set forth on page 3 are achieved in the PSI. By discussing the relative performance of the six groups of subjects we hope to gain some idea of the dependence of PSI performance on the level of difficulty of the problem, on familiarity with PSI problems, on the level of education of the subject, on the specialized skills of the subject as indicated by technical training, and on the interests of the subject as indicated by career plans.

In a few cases we will indicate cor-

relations with other variables. These correlations will refer only to problem 2, unless otherwise stated. The Examiner's Office of the University of Chicago has permitted us access to the results of various tests administered to students by that office. Among the tests administered were the American Council on Education Psychological Examination and a mathematical aptitude test. In addition, a number of tests were administered which have subtests designated as "analytic" scales. An average score for the combined analytic sections of these tests was computed and will be referred to as Total Analysis. Some of the correlations between these tests and PSI variables will be presented in what follows.

Our concern here, then, is to understand the set of PSI variables, rather than the differences between our six groups.

E. Work Variables

1. *TIME*—the number of minutes required for solution to be achieved. Solution criterion is the ability to light the center light using only buttons 4, 5, and 6 in some combination or sequence, *on three successive attempts*. (Data for this variable are presented in Table 1.)

Correlations. Time correlates with Effort (see below) .88, with the ACE "T" scale $-.36$, with the ACE "Q" scale $-.37$, and with Total Analysis as described in the preceding section $-.40$.

Discussion. Time is a power index rather than a process index. Certain relationships between Time and process are, of course, to be expected, since the process must occupy time. Inspection of the data shows that the various groups do not differ markedly on the time which they require for solution of problem 1. The difference is appreciably greater on problem 2. Note that groups 2, 5, and 6, which are the college student groups, require less time than groups 1, 3, and 4, which are the Ph.D. groups. As will be seen later, the faster achievement of solution by these

TABLE 1
TIME

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. college)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	23.8	22.0	18.4	22.0	23.0	20.7	21.4
Problem 2	44.5	41.0	31.0	39.2	41.8	37.2	50.0
Difference	-20.7	-19.0	-12.6	-17.2	-18.8	-16.5	-28.6

groups is accompanied by what might be termed lesser economy of inquiry. When the difficulty of the problem is increased, the same differential is manifested with the exception of group 6, the non-natural-scientist college group, which encounters real difficulty with various aspects of this problem. The increase in required time as the difficulty is increased is approximately the same for all groups except group 6. Thus it appears that the effect of education as such on the time required for solution of a PSI problem is to slow the process. The effect of special technical interest or training, as seen in the differential between groups 3 and 4 and groups 5 and 6, appears to cause slightly faster performance (where "technical" is used to denote natural science skills). None of these differences seems particularly large. The over-all correlation between performance on the two problems for this variable is .46. This correlation is not really a reliability index since the two problems are not equivalent in their structure. The increased familiarity of the subjects with the PSI after solving problem 1 does not bring about faster performance on problem 2. An accurate measure of the interaction would require varying the order of administration, which has not yet been done.

2. QUESTIONS—the number of questions required for solution to be achieved.

A question is defined as the series of manipulations of the PSI which occurs between the activation of any element in the net and the first subsequent time period in which all ele-

ments are inactive. In other words, a question is considered to be a manipulation or series of manipulations which interact or are contemporaneous. Both the specific content of each question and the time at which it was asked are recorded. (Data for this variable are presented in Table 2.)

Correlations. Questions (Q) is related to Effort (see following) .97, to Rate (see following) .49, to Actual Redundancy (see following) on problem 1, .72 and on problem 2, .77, to the ACE "T" scale — .43, to the ACE "Q" scale — .35, and to Total Analysis — .46.

Discussion. The number of questions required for the solution to be achieved is also a power index. The nature of the questions, that is, the content of the questions, is a process index. The treatment of this data will be presented in the next section, on Information Variables. The correlation between Q and Effort will be easily understood when the reader realizes that Effort is approximately $\frac{1}{2}$ Q times T. The correlation between Q and Actual Redundancy is of great importance and will be discussed in detail in the presentation of that variable. As with time, the data suggest that there is a reasonably high relationship between the power aspects of PSI performance and ACE "T" and "Q" scales and that the relationship of the achievement of a product on the PSI to the ability involved in the "Analy-

TABLE 2
QUESTIONS

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	69.1	35.0	51.0	23.0	46.0	57.8	77.2
Problem 2	110.0	50.0	77.0	41.0	59.0	82.3	156.0
Difference	-40.9	-15.0	-26.0	-18.0	-13.0	-24.5	-78.8

Bureau Ednl. Res. Research
DAVID H. ... COLLEGE

Dated

TABLE 3
COMPLEXITY

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	1.71	1.67	1.80	1.53	1.76	1.78	1.72
Problem 2	2.30	2.08	2.70	1.98	2.15	2.03	2.55
Difference	-.59	-.41	-.90	-.45	-.39	-.25	-.83

sis" subtests mentioned above is definite, being significant at the .01 level. The correlation for all subjects between performance on problems 1 and 2 is .46, the same as for Time. Note that groups 1, 3, and 4 require fewer questions than groups 2, 5, and 6 on both problems. This indicates that the more educated subjects ask fewer questions than the less educated ones. Note also that group 3 requires fewer Q than group 4, and group 5 requires fewer Q than group 6. Thus it appears that PSI performance using Q as an index is more economical as the educational level of the subject goes up, and, if the educational level is held constant, training or interest in the natural sciences is related to more economical performance. Finally, if we look at the increment in Q which accompanies an increment in the difficulty of the problem, we see that groups 1, 3, and 4 are affected less by this change than the other groups, although group 6 is much more troubled by this increase in difficulty with respect to this index than any other group. This index then is definitely not independent of education, or specialized training or interests. Furthermore, the increase in familiarity from the first to the second problem is not enough to offset the apparent slight increase in complexity.

3. COMPLEXITY—the total number of manipulations required for solution to be achieved, divided by the total number of questions (Q). This is the average complexity over the process as a whole. Complexity is also computed question by question (discussed under Analytic-Synthetic Shift below). (Data for this variable are presented in Table 3.)

Correlations. None computed.

Discussion. Complexity is an index which has more of the aspects of a process measure than the two indices so far discussed. Characteristically, during a performance questions require longer and longer series of manipulations, but are generally single manipulations at the beginning of the performance. The differential com-

plexity of different phases of the process is used as an index of the mode of approach as the process evolves (see section on Approach Variables following). Here we are dealing with overall complexity, which is an average value retaining little of the information inherent in the measure itself as a process index. Examining the data we see that on the simpler problem there is little variation in the mean complexity from group to group. As the difficulty of the problem increases, however, we see that the groups differentiate, with groups 2 and 6 showing an appreciably greater increase in complexity than the other groups. The differential between groups 1 and 2 and groups 5 and 6 suggests that the greater the educational level the less the complexity (1 vs. 2), and the more familiarity or interest in the natural sciences the less the complexity (5 vs. 6). The extent to which the factors of educational level and specialized training or interest affect this index appears to depend on the severity of the demands made on the individual by the problem. Increased familiarity with the PSI does not cause a decrease in the complexity index in this case.

4. RATE—the number of questions required for solution divided by the Time. (Data for this variable are presented in Table 4).

Correlations. Rate correlates with Effort (see following) .37, with Actual Redundancy .57, with Questions .49, with Pauses (see following) -.80, with Mixture of Modes (see following) .08, with Predominant Mode (see following) -.18.

Discussion. Rate is the first PSI variable we have discussed so far which is a process variable. Examination of the data for this index shows clearly that there is an underlying integrity and unity to the process of problem-solving which culminates in the achievement of a product (correlation coefficient between problem 1 and problem 2 is $r = .84$), yet varies greatly from individual to individual.

TABLE 4
RATE

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	3.01	2.02	3.42	1.77	2.25	2.62	4.03
Problem 2	2.27	1.47	2.62	1.17	1.76	1.94	3.10
Difference	.74	.55	.80	.60	.49	.68	.93

If one plots a graph of cumulative questions versus time, which we will subsequently term the Output Graph, the curve so generated is characteristically a straight line, the slope of which approximates Rate. The regularity with which such a straight line is observed suggests that in the PSI situation an individual evaluates the information which he possesses in such a way as to come to a decision in a characteristic time. While on the whole the performances can be

approximated by a straight line, deviations from this straightness occur in two ways. Occasionally one observes inflection points at which a change in slope occurs (see Inflection Points following), and occasionally one observes pauses in output followed by a transient period of accelerated output until the curve returns to the basic straight line, after which the previous slope is resumed (see Pauses following). In Fig. 4, a number of Output Graphs are presented.

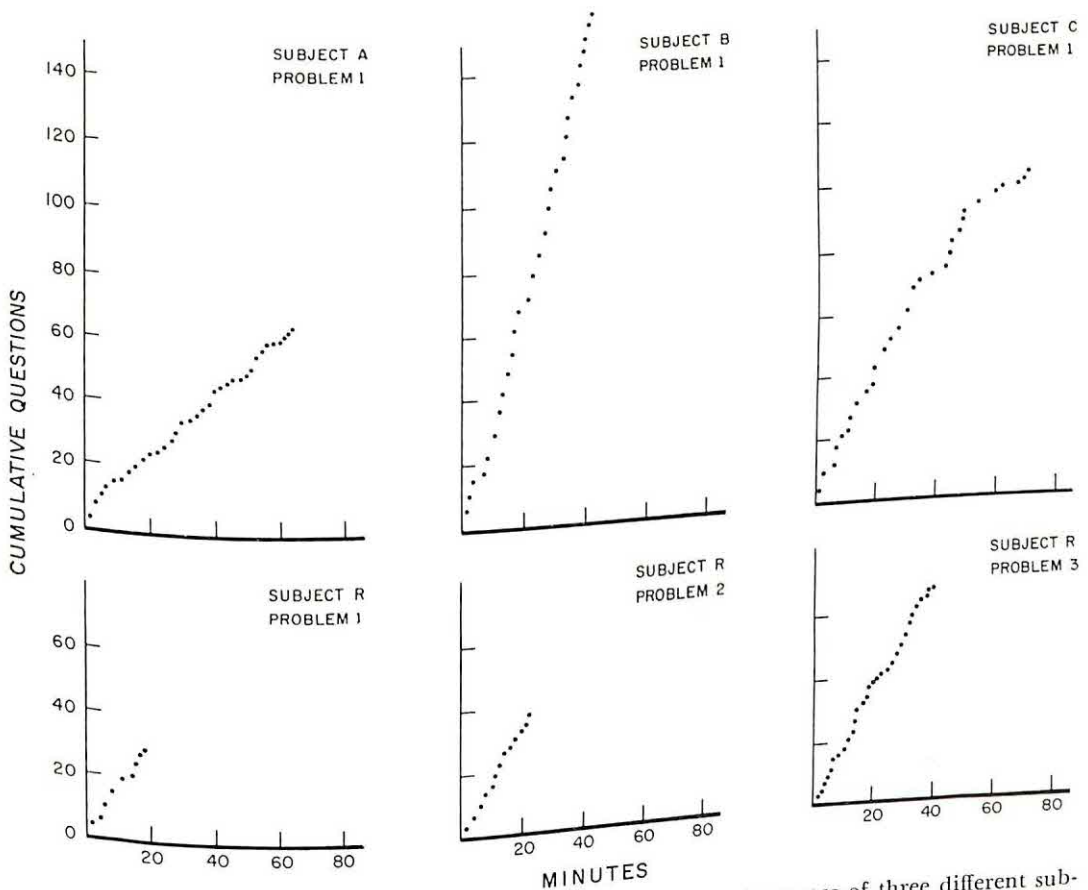


FIG. 4: Output Graphs. Note: The top three graphs represent performance of three different subjects on the same problem. Notice the similarity in slope of the three bottom graphs, representing performance of the same subject on three problems of increasing difficulty.

The three top graphs in Fig. 4 represent three different persons solving the same problem. The top left curve represents an average performance; while it is approximately a straight line, there are a number of minor deviations from linearity. The top middle graph represents a highly linear performance, obtained fairly often (see Changes of Set following). The top right graph represents an unusual performance with multiple inflection points. The bottom three graphs represent the performance of the same individual on three problems of increasing difficulty. Note the similarity in slope despite the increase in difficulty.

From the correlations above we see that as Rate increases, the economy of Effort and Questions increases, the inefficiency of handling information (Actual Redundancy) increases and the frequency of Pauses decreases. The relation of Rate to the two Mode indices shows that speed of asking questions is essentially independent of whether the subject is in the Analytic or Synthetic phase of the process (see discussion under Approach Variables below).

The data indicate that both level of education and specialized technical training or interest are related to performance at lower rates. A general slowing accompanies the increase in difficulty, but is markedly differential.

5. **PAUSES**—the number of minutes during which no questions were asked, divided by the total Time required for solution. (Data for this variable are presented in Table 5.)

Correlations. Pauses correlates with Effort (see below), $-.22$, and with Rate $-.80$.

Discussion. Pauses is an index which gives us an insight into the constancy with which the process goes on. It is possible to have two performances which display the same Rate, one of which has no pauses and the other of which contains an appreciable number of pauses, because of the smoothing which enters into the computation of Rate. In the two cases the distribution

of actions differs appreciably with respect to grouping. As can be seen from the correlation of $-.80$ between Pauses and Rate, the relationship between these two is high, yet the evenness of spacing of questions is not the determinant of Rate. Note the large differentials in the above data. We see that educational level seems to affect the frequency of pauses, by contrasting the data for groups 1, 3, and 4 with those for groups 2, 5, and 6. By comparing group 3 with group 4, and group 5 with group 6, we see that at the same educational level the group with "technical" training or interest displays a higher incidence of pauses. This holds true at both levels of difficulty. There is a general increase in the frequency of pauses as the problem increases in difficulty. Note that the higher the initial frequency of pauses, the greater the increase in frequency which accompanies greater difficulty. Thus this index appears to be somewhat dependent on both the educational level and the special skills of subjects. Increased familiarity with the PSI does not result in a decrease in frequency of pauses. The correlation between the two problems for this index is $.76$.

6. **INFLECTION POINTS**—the number of inflection points in the Output Graph. (Data for this variable are presented in Table 6.)

Correlations. Inflection Points correlate with Effort (see following) $.42$.

Discussion. Inflection Points are of utility in designating points in the process at which there has been a change in the rate at which the process is being generated. Since we interpret the Rate as an index of decision-making time, a change in the slope of the output graph represents a change in decision-making time. We believe that the constant Rate which is manifested by the straight line form of the Output Graph indicates a *constant set* on the part of the subject. In support of this interpretation are the following facts: As will be described in the section of this paper on Information Variables, one can stipulate the point in the process at which

TABLE 5
PAUSES

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.09	.20	.05	.26	.15	.12	.03
Problem 2	.18	.30	.11	.39	.23	.20	.06
Difference	-.09	-.10	-.06	-.13	-.08	-.08	-.03

TABLE 7
PERCENTAGE OF NONLINEARITY

Problem	All Groups
Problem 1	12.6%
Problem 2	22.5%

of the performance after the third minute.) The area of the large triangle under this straight line is determined, as is the area between the extended straight line and the actual performance polygon, indicated as the small triangle labeled "deviation from linearity" in Fig. 5. The ratio of the latter to the former is percentage of Nonlinearity. (Data for this variable are presented in Table 7.)

Correlations. Percentage of Nonlinearity correlates with Actual Redundancy (see following) .24.

Discussion. This index is used in conjunction with the previous one in order to estimate the extent to which an inflection point affects the subsequent process. We argue that if a change in set is major, the characteristic decision-making time which appears to be related to set or conceptual framework (see previous discussion) should change appreciably. This index measures the accumulated amount of change in decision-making time summed over the entire process from the point of change and stated as a percentage. We assume that if there were no change in slope of the output graph, the first straight line, extended to the time of solution, would closely approximate the actual performance. To the extent that this performance proceeds at a rate different from this initial rate, the area between the performance polygon and the extended initial slope will increase. By division by the total area under the extended initial slope, we state this deviation as a percentage, the sign

of which indicates whether the change in rate was an increase or decrease.

Percentage of Nonlinearity tends to be larger on the more difficult problem. Also, as the % NL goes up, the performance tends to become somewhat more inefficient insofar as utilization of information is concerned. The correlation between performance on this variable from problem 1 to problem 2 is .02. Further investigation with parallel forms of PSI problems, equated for difficulty, will be necessary before one can determine whether the low correlation indicates unreliability of this index, or results from the generally less efficient performance which subjects display on the more difficult problem.

8. EFFORT—the area under the performance polygon on the Output Graph.⁵ (Data for this variable are presented in Table 8.)

Correlations. Effort correlates with Pauses -.22, with Questions .97, with Rate .37, with Actual Redundancy (see following) .72, with Time .88, with Inflection Points .42, with Changes of Approach (see following) .59, with Relative Analytic-Synthetic Shift -.59, with Absolute Inferential Lag .77, with Relative Inferential Lag .30, with ACE "T" scale -.45, with ACE "Q" scale -.33, with ACE "L" scale -.36, and with Total Analysis -.51.

Discussion. Effort was adopted as an index to attempt to incorporate both Time and Questions in one measure. As can be seen from the above correlations, these two measures are sufficiently highly related so that Effort does not tell us much that either of the simpler measures could

⁵ Effort was measured by plotting the Output Graphs on standard axes and measuring the pertinent areas by use of an Ott Planimeter.

TABLE 8
EFFORT

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	379.	155.	299.	67.	235.	298.	385.
Problem 2	815.	408.	462.	321.	488.	568.	1095.
Difference	-436.	-253.	-163.	-254.	-253.	-270.	-710.

not do as well. Yet, because it did combine the two into a single index, a number of correlations were computed using this index. They can be summarized as follows: As effort increases, pauses tend to decrease, rate tends to increase, changes of set or inflection points become more frequent, shift from analysis to synthesis and back to analysis occurs more often, inferences are less validly drawn and tested, and the shift from analysis to synthesis tends to occur relatively earlier in the process. The correlation between performance on the two problems for this measure is .42.

Examination of the data above indicates that this measure is dependent both on educational level and special technical training, as must be expected to be consistent with the dependence of the two components of this index on the same factors. Note the surprising uniformity of the increment in Effort with increased difficulty of the problem for the various groups except group 6.

Note that both ACE "Q" and "L" scales correlate significantly at the .05 level with Effort, which indicates that verbal as well as quantitative skills are useful in PSI performance, yet clearly neither is sufficient to be the major determinant of performance quality. Finally, the high correlation, significant at the .01 level, with Total Analysis indicates that the skills involved in PSI performance are similar to the skills involved in the elucidation and application of principles in other areas more relevant to real-life problems.

F. Information Variables

At each step in the problem-solving process, as we observe it in this special situation, two kinds of information are obtained by the subject:

1. Information overtly explicit in the answer to the questions asked of the machine, referred to as Q_e .
2. Information obtainable by logical inference from the answer to the question asked, taking into consideration the implications of all previously acquired knowledge. This will be referred to as Q_i .

We can represent the process of acquisition of information from the apparatus by an individual as the sum yield of a series of questions, each of which contains in its answer an explicit (Q_e) and an implicit (Q_i) amount of information. It is clear that under certain con-

ditions Q_i will be a function not only of the question just answered, but of information previously obtained. Thus, *the total information content of the answer to a particular question may not be an invariant characteristic of the question, but a function of the sequence of questions asked previously.* The process then can be represented as follows:

Total information available after the n th question =

$$\Sigma(Q_e + Q_i)_1 + (Q_e + Q_i)_2 + \dots +$$

$$(Q_e + Q_i)_{n-1} + (Q_e + Q_i)_n = \sum_{j=1}^n (Q_e + Q_i)_j =$$

the sum of all explicit and inferred information.

If the *explicit* information content of the answer to any question, $(Q_e)_n$, is contained in the sum of the explicit and implicit information content from the first answer to the n -first answer elicited from the apparatus in the process, we will term that question *inferable*. Obtaining the answer to that question will not further increase the sum of elicited information. It is possible, by changing the ordering of a given set of questions, to vary greatly the number of inferable questions in the set. If we define the *actual redundancy* as the number of inferable questions in a particularly ordered set divided by the total number of questions in that set, it is possible to show that the actual redundancy of a particularly ordered set can be appreciably less than the actual redundancy of that set differently ordered. Similarly, it can be shown that the redundancy of a particularly ordered set can be less than the redundancy of another set which contains fewer questions.

We thus reach the conclusion that if we wish to quantify the efficiency of acquisition of information in a given problem-solving process, it is not adequate merely to consider the number and

choice of questions asked in the process, treating the information content of the answer as only a function of the particular question. In order to state the efficiency of a given process, that is, in order to state the extent to which the information which was acquired up to a given step was used to determine the decision as to what should be the next step in the process, we must recognize that the rate of acquisition of information may well depend on the *ordering* of questions in addition to the *nature* of the questions.

Further, we reach the conclusion that there may be a fortuitous factor involved to some extent in the problem-solving process. Although there may be no a priori basis for asking a particular question first in a series, yet the selection of that question as a starting point, or the asking of that question before some other question but after others, will affect the extent to which portions of the information content of the total pool can be inferred at any point in the subsequent process. (Note: This would appear to have interesting consequences in a number of areas. It would seem that to score an examination or a psychological test, in which the contribution of this kind of factor to the performance has *not* been ruled out, on the basis of the amount of time taken or any similar nonsequential "power" criterion for quality of performance, is to include the possibility that a subject with a poor score is actually more efficient than a subject with a "good" score. Consideration of the residual variance in actual redundancy other than that due to the number of questions [discussed below] shows that *ordering is as important as what is asked.*)

In order to partial out the contribution of this fortuitous factor of initial

choice, and in order to enable the precise evaluation of redundancy by accounting for the dependence of information content on ordering, we have analyzed each problem used in this study in the following way: A flow sheet has been developed which shows, *for any possible series of questions*, what the explicit and implicit information obtained at each step will be. The necessary and sufficient information for the solution of each problem and the total content of the information pool for each problem is indicated on each flow sheet. Detailed exposition of such a flow sheet for a typical problem requires too much space to be here presented. Suffice it to say that when such a flow sheet has been constructed, a typical problem-solving performance can be sequentially analyzed for the efficiency of information handling in a few minutes, with absolute reproducibility and precision.

The analysis also indicates the point in the performance where the necessary and sufficient items of information for the solution of the problems have been *explicitly* achieved, subsequently referred to as the *NASI(exp)*, and the point where these could be implicitly achieved if maximum inference were validly drawn, subsequently referred to as the *NASI(imp)*.

Other measures of redundancy can be defined. There are a number of different kinds of redundancy, for example, the number of times in a particular performance that the *same* question is repeated. We have selected this particular index as best suited to our present purpose, but recognize that redundancy can meaningfully be defined in a number of alternative ways, and call to the attention of the reader such potentially useful measures of redundancy as repetition of particular questions or patterns of ques-

TABLE 9
EXHAUSTIVENESS OF INQUIRY

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	13.6	12.4	13.5	11.8	12.9	13.3	13.9
Problem 2	13.0	12.8	13.4	12.1	13.4	13.5	12.4
Difference	.6	-.4	.1	-.3	-.5	-.2	1.5

tions, which are frequently encountered.

We are now in a position to define a number of different variables related to the acquisition and handling of information:

1. **EXHAUSTIVENESS OF INQUIRY**—the number of different items actually elicited from the information pool. There are, in each of the two problems herein discussed, only 17 possible different meaningful items. All questions can be decomposed into aggregates of these basic items. (Data for this variable are presented in Table 9.)

Correlations. Exhaustiveness of Inquiry correlates with Actual Redundancy (see following) .10.

Discussion. This variable tells us the extent to which the set of possible "basic" questions has been exhausted by the subject. Given two subjects with the same number of questions required to solution, this index is a measure of the stereotypy of response. If one analyzes the breakdown of total questions into separate items, one finds a marked inhomogeneity of distribution. Thus, of 174 questions asked by one subject in a performance on the first problem, 85 consisted of or included the same item. It might be of interest to plot the rank order of the frequency with which a given item appears versus its actual frequency in order to see whether the logarithmic relation described by Zipf for word frequency distributions can be demonstrated with a symbolic language of this abstract type. One might also tabulate response frequencies with the end of categorizing "popular" and "original" responses. These analyses have not been performed.

A frequency analysis of the sort described above, were the logarithmic prediction upheld, would give us insight into the mode of organization of the process in a descriptive sense. As this

variable is used at present, it is more a product than a process index. Examination of the correlation between this measure and Actual Redundancy (see following), which is a process index, shows that the two are almost completely independent. This is an unexpected finding. One would normally expect to find that as stereotypy of response increases, redundancy increases. This apparent contradiction is probably due to the existence of two sources of low values for this measure: first, the incisive performance requires only a few items to be elicited; second, the disorganized performance may be highly stereotyped, or may be highly exploratory but undirected. Given two persons who required the same number of questions for solution, one might expect the one with a lower Exhaustiveness measure to demonstrate less flexibility conceptually, provided that both performances were relatively long. The data suggest that educational level and specialized training affect this variable very slightly. It is of interest that members of all the groups consistently select two items in each problem which can be inferred by inspection a priori.

2. **ACTUAL REDUNDANCY**—the total number of inferrable questions, as defined in the preceding discussion, divided by the total number of questions asked. (Data for this variable are presented in Table 10.)

Correlations. Actual Redundancy correlates with Questions (problem 1) .72, (problem 2) .77, with Rate .57, with Effort .72, with Exhaustiveness of Inquiry .10, with Percentage of Nonlinearity .24, with Mixture of Modes (see following) -.24, with Predominant Mode (see following) .08, with Frequency of Change of Approach (see following) .49, with Relative A-S Shift (see following) -.39, with Absolute Inferential Lag

(see following) .59, with Absolute Synthetic Lag (see following) Explicit $-.41$ and Implicit $-.47$, with ACE "Q" scale $-.24$, ACE "L" scale $-.14$, and Total Analysis $-.46$.

Discussion: Actual Redundancy is a process variable, in the sense that it describes the logical efficiency of the problem-solving performance as a whole by examining it step by step. As pointed out in the theoretical discussion in which this measure was developed (see Section F, paragraph 5), one expects that ordering as well as the actual questions asked will determine the rate of acquisition of information. In support of this, note that the correlation between Questions and this variable is .72 and .77 for the two problems used. This indicates that about 48% of the variance in Actual Redundancy on the easier problem, and 41% on the more difficult problem, must be attributed to the *ordering* of the questions asked. This gives some idea of the importance of sequential analysis of the sort here used in proper evaluation of the quality of a process.

Further inspection of the correlations presented above suggests that fast workers tend to be more redundant. Redundancy of this sort is slightly related to a tendency to carry out analysis separate from synthesis, as we see from the correlation with Mixture of Modes. It is relatively independent of whether the performance is predominantly analytic or synthetic, as we see from the correlation of .08 with Predominant Mode. The correlation with Frequency of Change of Approach shows that Redundancy is related to a large amount of shift from synthesis back to analysis, and vice versa. As Redundancy goes up, the change from the analytic phase to the synthetic phase comes relatively earlier in the performance; this change occurs more prematurely in the less redundant workers, with respect to the achievement of the information needed for solution, as we see from the correlations with Relative A-S Shift and Absolute Synthetic Lag. Finally, as Redundancy increases, the lag in inference increases, as seen from the correlation with Absolute Inferential Lag.

The correlation of this measure from problem 1 to problem 2 is .60.

Examination of the data above shows that this index is somewhat dependent on educational level and specialized training or interests of a technical sort. The first effect may be seen by comparing the more educated groups (1, 3, and 4) with the other groups. Note, however, that group 5, which is a college group, is as effective as group 4, which is a Ph.D. group, with respect to this index. Note the relatively uniform increment in Redundancy which all groups display as the difficulty of the problem increases.

Note also that the correlation of this index with the ACE scales is not high, while there is a correlation significant at the .01 confidence level between this index and the derived score for Total Analysis on all achievement tests for which data are available.

3. NASI POINT (IMPLICIT)—the point in the performance where the subject obtains the information which is necessary and sufficient to enable solution of the problem, provided that maximum inferential use is validly made of the data on hand. This point is stated in two ways: (a) Absolute—the number of questions asked up to and including the point; and (b) Relative—as a percentage of the total number of questions. (Data for these variables are presented in Table 11.)

Correlations. None computed.

Discussion. This measure serves as an index of the end of the first phase of inquiry in the problem-solving process as observed using the PSI. Once the NASI (Implicit) has been achieved by the subject, he is theoretically able to solve the problem. That is, he has had the opportunity to learn the required relationships on which a solution must be based. Note that for all groups this point is achieved at about the same number of questions in both problems, with the excep-

TABLE 10
ACTUAL REDUNDANCY

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.79	.70	.83	.61	.79	.79	.89
Problem 2	.86	.78	.92	.69	.86	.83	.94
Difference	-.07	-.08	-.09	-.08	-.07	-.04	-.05

TABLE 11
NASI POINT (IMPLICIT)

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	14.6	14.2	17.9	12.2	16.1	12.2	12.8
Problem 2	15.4	12.7	18.3	10.6	14.6	11.4	20.3
Difference	-.8	1.5	-.4	1.6	1.5	.8	-7.5
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.40	.47	.41	.56	.39	.41	.24
Problem 2	.27	.34	.26	.40	.27	.26	.18
Difference	.13	.13	.15	.16	.12	.15	.06

tion of group 6. The slight increase in difficulty involved when the second problem is presented has little consequence on the location of the NASI (Implicit), Absolute. There is evidence that educational level or special training are effective in slightly earlier achievement of the NASI (Implicit).

The NASI (Implicit) Relative tells us the percentage of the total process required for the achievement of the NASI (Implicit). Note that as educational level or special technical training increases, the relative location of the NASI (Implicit) is later in the performance. Since we know from the absolute location that there is little difference in the numerator of the fraction which determines the relative value, the difference which we observe here is mostly due to the denominator, i.e., total questions in the various performances. Note that the effect of the increase in difficulty is to shift the location of the

NASI (Implicit) point about 13% earlier in the performance, except for group 6.

4. NASI POINT (EXPLICIT)—the point in the performance where the subject obtains the information which is necessary and sufficient to enable solution of the problem without the necessity of inferring any relationships involved. (That is, the relevant items have been overtly elicited from the PSI.) This point is stated in two ways: (a) Absolute—the number of questions asked up to and including the point; and (b) Relative—as a percentage of the total number of questions. (Data for these variables are presented in Table 12.)

TABLE 12
NASI POINT (EXPLICIT)

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	46	31	53	21	40	53	59
Problem 2	66	38	71	24	50	68	110
Difference	-20	-7	-18	-3	-10	-15	-51
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.87	.02	.84	.94	.90	.88	.80
Problem 2	.74	.81	.68	.77	.86	.73	.72
Difference	.13	.11	.16	.17	.04	.15	.08

Correlations: None computed.

Discussion. This measure serves as an index of the end of the second phase of inquiry in the problem-solving process as observed using the PSI. We conceive of the first phase, ending with NASI (Implicit), as an interval in which basic structuring of the problem occurs. At the end of the first phase, the individual has elicited the information which is required minimally for solution to be achieved. Almost invariably, solution based on such inferences is not forthcoming. Instead, there intervenes a period of further inquiry which culminates in the NASI (Explicit), in which the material which might have been inferred at the end of the first phase is overtly elicited. This second phase we tend to conceptualize as a reassurance period in which verification takes place of the inferences which can be drawn after the first phase. Of course, the fact that inferences *can* be drawn does not mean that the subject actually *does* draw them in the second phase. It might well be that the activity in the second phase is frequently required for perception of relationships to become accurate. Yet it is of interest that one frequently observes subjects who verbalize as they work in the second phase, and it becomes apparent that inferences are being tested. Further, very seldom does one observe a performance in which the subject does *not* go ahead after achieving the NASI (Implicit) to achieve it explicitly. Subjects who change to synthetic behavior, as defined later, before the NASI (Explicit) is achieved, almost invariably revert to analysis subsequently until the NASI is achieved explicitly.

Examination of the data shows that a marked differential exists on the absolute index between our various groups. This differential is such as to support the interpretation that both educational level and specialized technical training or interests contribute to the ease with which the subject achieves the NASI explicitly. Since the various groups end the first phase with the implicit achievement of the NASI almost identically, this difference in the location of the NASI (Explicit) point must be due to the variable length of the interval *between* the two points. This will be discussed in the next section.

Note that groups with higher educational levels and greater interest or training in specialized technical areas tend to achieve the NASI explicitly relatively later in the over-all performance. This implies that these groups have less difficulty in combining and applying the results of analysis than groups with less education or technical sophistication.

Finally, note that the decrement in the Relative NASI (Explicit) as the difficulty of the problem is increased is almost identical with that for NASI (Implicit), with the exception of group 4.

This means that as the difficulty of the problem is increased there is a decrease in the portion of the problem which is occupied by the first and second phases of analysis. Therefore it follows that there must be a compensatory increase in the relative portion of the process occupied by activities other than analysis of this sort, as difficulty increases. In other words, the effect of increasing the difficulty of the problem is not to cause relatively more analytic trouble, but rather to cause relatively more synthetic difficulty.

5. INFERENCE LAG—the number of questions which intervene between the achievement of the NASI implicitly and explicitly. The Inferential Lag can be stated in two ways: (a) Absolute—the actual number of questions in the interval between the two NASI points; and (b) Relative—the number of questions in the interval divided by the total number of questions required for solution. (Data for these variables are presented in Table 13.)

Correlations: Absolute Inferential Lag correlates with Effort .77 and with Actual Redundancy .59. Relative Inferential Lag correlates with Effort .30.

Discussion. The two Inferential Lag indices are measures of the quality of the inferences drawn by the subject in the course of the process. They are indicative of the failure of the subject to infer properly from the data he has gathered or of his need for reassurance about inferences which he has made, which he tests by making the conclusion overt rather than by attempting to utilize the conclusions in attempted solutions. The fact that the NASI (Explicit) occurs in a performance once the NASI have been achieved implicitly is evidence that conclusions which could have been drawn either have not been drawn or have not been accepted. In the interval, the subject accumulates direct evidence which provides the NASI without inference. Since this material is redundant, it is obvious that this index should correlate well with our redundancy index for the whole performance, as it does. It is also obvious that difficulty in the second phase of the process, as measured by the lag, will correlate significantly with our index of over-all difficulty—Effort.

Examination of the data for the absolute index suggests strongly that both educational level and the amount of training or interest in technical fields are related to the size of the inferential

TABLE 13
INFERENCEAL LAG

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	40.0	17.2	34.7	9.8	24.0	40.7	46.3
Problem 2	62.7	25.5	34.0	13.7	36.5	55.5	89.0
Difference	-22.7	-8.3	0.7	-3.9	-12.5	-14.8	-43.7
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.52	.45	.44	.38	.50	.46	.56
Problem 2	.50	.48	.40	.37	.50	.48	.54
Difference	.02	-.03	.04	.01	-.09	-.02	.02

lag. Note, however, that the consequence of increasing difficulty of the problem is to lengthen this interval for all groups except the first-year college group, group 2. The difference between groups 1, 3, and 4—the more educated groups—and the less educated groups is more perceptible on the easier problem. The difference between groups 3 and 5 and their less technically sophisticated counterparts at equivalent educational levels tends to be more marked on the more difficult problem. This suggests that as the complexity of the task increases, education as such is less advantageous than technical sophistication.

Note that the more technically sophisticated groups devote relatively less of the over-all PSI performance to the inferential lag, while this distinction does not appear in favor of the more educated group compared with the first-year group (1 vs. 3). Finally, let us point out that the effect of increasing the difficulty of the task on the proportion of the over-all performance spent in the inferential lag is remarkably little. This suggests that increase in difficulty is not followed by an increase in the relative difficulty of inference. However, the correlations between performance on the two problems are .30 for the absolute index and $-.12$ for the relative index. The first correlation is significant at the .05 level. This indicates that the effect of increasing the difficulty of the problem is not the same on all individuals. Again, it should be pointed out that the correlations just given do not indicate reliability, since this is a measure of organization, and one has no reason to expect similar organization of process for two tasks which differ in the severity of the demands made on the individual. Tasks of equated difficulty must be used to obtain such a measure of reliability and this has not been done.

G. Approach Variables

A graph is drawn of the duration of the average question in each minute plotted versus time, to which we will subsequently refer as the Analysis-Synthesis (A-S) Graph (see Section I). To draw this graph, we sum the series of manipulations of the apparatus which occur in each minute, divide by the number of questions asked in that minute, and plot the result versus that minute.

We assume that the following distinction can be made between two modes of approach to the problem. Direct questions about particular propositions or relationships can be asked with very short series of manipulations. Such short manipulative combinations we term as of the *analytic* mode. When the subject attempts to utilize relationships which he has analyzed to achieve solution of the problem, combining them and causing interactions to occur between them, the series of manipulations composing a question becomes longer. Such long manipulative combinations we term the *synthetic* mode.

Rigorously, the A-S graph is analyzed as follows: A time interval is classed as *analytic* if there are less than two ordered

related manipulations (steps) per average question in that interval, *or* if there are exactly two steps per question in that interval and the interval is bracketed on either side by intervals with less than two steps per question, *or* if there are exactly two steps per question in that interval and only one adjacent interval has less than two steps per question and only one question was asked in the relevant interval. A time interval is classed as *synthetic* if there are more than two steps per question in that interval, *or* if there are exactly two steps per question in the interval and it is bracketed by two intervals each of which contains more than two steps per question, *or* if there are exactly two steps per question and only one adjacent interval has more than two steps per question and more than one question was asked in the relevant interval. Note that this analysis could be carried out question-by-question instead of minute-by-minute. Such an analysis would be more precise but much more laborious and has not been extensively pursued.

The performance can in this fashion be characterized as a period of so many minutes of analysis (A) followed by a period of so many minutes of synthesis (S) followed by . . . etc. Consider a performance in which five minutes of A is followed by three minutes of S is followed by five minutes of A is followed by seven minutes of S. We write this as 5A3S5A7S. The total time of this performance is 5 plus 3 plus 5 plus 7, or 20 minutes. There was a total of 10 minutes spent in the analytic mode, or 50%. Ten minutes were spent in the synthetic mode, or 50%. (Note that if this analysis were done on the basis of a question-by-question analysis, the straight line phenomenon described in the section on Output Graphs would result in approxi-

mately the same distribution of analysis and synthesis.)

It can be shown that, at any point in a performance, the following is true: the percentage of analytic before that point *minus* the percentage of synthetic before that point *is equal* to the percentage of synthetic after that point *minus* the percentage of analytic after that point. We are now in a position to define a number of variables.

1. ANALYTIC-SYNTHETIC SHIFT POINT (A-S SHIFT)—that point in the performance where the percentage of Analytic before *minus* the percentage of Synthetic before is a maximum for the whole performance. This defines a unique point which separates the PSI performance into a predominantly analytic phase before the point and a predominantly synthetic phase after the point. This index can be stated as (a) Absolute, and (b) Relative. (Data for these variables are presented in Table 14.)

Correlations. The Relative A-S Shift correlates with Effort $-.59$, and with Actual Redundancy $-.39$.

Discussion. The Analytic-Synthetic Shift Point, together with the indices to be presented in the remainder of the section on approach variables, permits us to measure the extent to which the PSI performance does consist of separate and separable phases of information gathering and information application. It will be seen in what follows that such a separation is possible, and that examination of the relationship between the state of information possessed by the subject and the transition from inquiry to application is quite informative. Note that as the shift comes *relatively* earlier in the performance, both Effort and Redundancy tend to increase.

Examination of the data above suggests that more educated groups tend to shift earlier than less educated groups, and that more technically sophisticated groups tend to shift earlier than less sophisticated groups. Note that these differences do not hold for the relative location of the shift point. It is exceedingly interesting that the change in the absolute shift point is so little when the difficulty of the problem is increased.

TABLE 14
ANALYTIC-SYNTHETIC SHIFT POINT (A-S SHIFT)

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	46.3	26.3	49.0	18.4	33.7	28.0	62.5
Problem 2	47.9	28.4	48.5	24.2	32.3	35.0	66.5
Difference	-1.6	-2.1	.5	-5.8	1.4	-7.0	-4.0
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.70	.75	.77	.79	.71	.65	.78
Problem 2	.55	.59	.73	.64	.54	.58	.50
Difference	.15	.16	.04	.15	.17	.07	.28

It suggests that after a certain amount of inquiry there is a tendency for the subject to shift into synthesis, "ready or not." However, we note from inspection of the data for the relative index that the shift occurs appreciably earlier for all groups on the second problem. From this we conclude that the consequence of increasing the difficulty of the task is to increase the proportion of the process which is in the synthetic phase, rather than to increase analytic aspects of the performance. That is, as the difficulty of the problem goes up, about the same effort appears to be required to acquire the necessary information but it becomes more difficult to reconcile the constraints which this information imposes.

The correlation between the two performances for the absolute index is .13, and for the relative index is .33. This suggests that there is a reorganization of the internal coherence of the PSI performance as difficulty increases.

2. PREDOMINANT MODE—total number of Analytic intervals divided by the total number of Synthetic intervals. (Data for this variable are presented in Table 15.)

Correlations. Predominant Mode cor-

relates .08 with Actual Redundancy and -.18 with Rate.

Discussion. This index is a measure of the relative amounts of analysis and synthesis of which a PSI performance is composed. We see from the correlation of .08 with Actual Redundancy that this index is relatively independent of the efficiency of the performance, in terms of economy of inquiry; the same holds true of Rate. The data show that on the easier problem there is little relation between educational level and the A-S ratio. The technically sophisticated groups tend to engage in somewhat less analysis relative to synthesis than their counterparts at the same educational levels. On the more difficult problem neither educational level nor special technical skill or interest appears to be related to the A-S ratio.

Notice that all groups except group 5 show a marked decrease in analysis relative to synthesis as the difficulty of the problem increases. This supports the inference which was earlier drawn that increased difficulty has the effect of causing greater synthetic effort rather than analytic effort. That is, the manifestation of increased difficulty in the PSI problems here used is to bring about a shift toward more synthetic en-

TABLE 15
PREDOMINANT MODE

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	6.3	5.6	5.0	5.0	6.2	4.1	9.7
Problem 2	3.3	1.7	1.8	1.7	1.7	3.9	5.3
Difference	3.0	3.9	3.2	3.3	4.5	.2	4.4

TABLE 16
 MIXTURE OF MODES

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.79	.85	.76	.91	.80	.75	.89
Problem 2	.66	.64	.81	.63	.65	.76	.53
Difference	.13	.21	-.05	.28	.15	-.01	.36

deavor in the process, because of the complexity of the constraints which must be reconciled. This should be considered in view of the fact that the two problems, as described earlier, do not differ in the total size of the pool of information items which stipulates the problems, but differ in the number of coincidences which must be achieved using these relationships. The reader may recall that the more difficult problem requires the subject to achieve four coincidences, the easier only three. Presumably, it is to this difference that we can attribute the above shift toward synthesis.

The correlation between the two performances for this index is .70. Thus, even though there is a shift in the internal organization of the process as the difficulty is increased, the way in which an individual distributes his efforts between the acquisition of information and the utilization of that information remains fairly characteristic, in the sense that one can relate the new organization to the old.

3. MIXTURE OF MODES—the percentage of the total number of analytic intervals in the performance which is located before the shift point minus the percentage of the total synthetic intervals in the performance located before the shift point. (Data for this variable are presented in Table 16.)

Correlations: Mixture of Modes correlates with Actual Redundancy $-.24$ and with Rate $.08$.

Discussion. This index measures the extent to which PSI performance can be separated into distinct analytic and synthetic phases. It is apparent from an inspection of the data that these two phases exist relatively separate from one another in both problems, although the definiteness of the separateness is greater on the more simple problem. Factors such as education or specialized technical training or interest do not appear to relate to the extent of this separation in a clear fashion. Note, however, that the definite separation decreases more over the more

educated groups than over their counterparts as the difficulty of the problem increases, with the exception of group 6. The meaning of this differential is not clear.

The low correlation of this index with Actual Redundancy and with Rate suggests that we have in the Approach Variables access to a set of dimensions relatively orthogonal to those which span the space of Work Variables and Information Variables.

In summary, then, this variable enables us to distinguish the intermingling of two subsidiary processes which are combined in PSI performance. The high values which are obtained, showing relatively clean separation of the process into two phases, indicates that the Analytic-Synthetic Shift Point, rather than being an artificial conceptual device, is a concept which has real functional utility, enabling the separation of two aspects of process which are not homogeneously distributed in the performance.

The correlation for this variable between the two performances is $-.39$. We note from the data above that the greater the separation between the two phases on the easier problem, the greater will be the changes as the difficulty of the problem is increased. Perhaps individuals whose performance on the easy problem is composed of exceedingly distinct phases are more venturesome on the second problem. Although the reason for this peculiar relationship is far from obvious, the correlation is significant at the .01 level (negative sign and all).

4. FREQUENCY OF CHANGE OF APPROACH—the total number of inversions from synthesis back to analysis which occur in the performance. (Data for this variable are presented in Table 17.)

Correlations. This index correlates with Effort $.59$, and with Actual Redundancy $.49$.

Discussion. The previous index, Mixture of Modes, gave us a measure of the separation of the two phases, but did not enable us to dis-

TABLE 17
FREQUENCY OF CHANGE OF APPROACH

Problem	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	3.6	2.7	3.3	2.0	3.3	3.4	2.6
Problem 2	7.8	7.0	5.7	6.5	7.4	4.8	10.8
Difference	-4.2	-5.3	-2.4	-4.5	-4.1	-1.4	-8.2

criminate between the performance which interrupted an analytic phase with a prolonged synthetic interjection and one which shifted back and forth repeatedly. The present index tells us which case we are dealing with. It measures the number of times that a subject attempts to synthesize and feels forced to revert to analysis for more information. We can see by comparing the correlations with Redundancy of this and the previous index that mixture of modes per se does not relate to a higher redundancy, but that frequent and premature attempts at synthesis as measured by frequency of change of approach do so relate.

Neither the frequency itself nor the change in frequency with greater difficulty appears to be highly related to either educational level or specialized technical training or interest. We note that all groups change more frequently on the more difficult problem.

The correlation between the two performances for this index is .04. This may be a consequence of the much greater spread for this variable on the more difficult problem.

5. SYNTHETIC LAG (IMPLICIT)—the Absolute A-S Shift Point minus the Abso-

lute NASI Point (Implicit). This index is stated as (a) Absolute and (b) Relative. (Data for these variables are presented in Table 18.)

Correlations. Absolute Synthetic Lag (Implicit) correlates with Actual Redundancy .47.

Discussion. This is a measure of the failure of the subject to shift from analysis to synthesis as soon as the sum of inferences which can be validly drawn from the information he has gathered is sufficient for solution to occur. Negative implicit synthetic lag occurs very rarely.

The most impressive fact about this index is its remarkably small change when the difficulty of the problem is increased. Neither the absolute nor the relative versions of this index show large changes with difficulty, except for group 6 in the relative category. These data show that shift tends to occur after a characteristic interval, independent of the objective difficulty presented by the problem, if one considers group averages. This interval is shorter for the educated groups than for the less edu-

TABLE 18
SYNTHETIC LAG (IMPLICIT)

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	29.0	12.6	31.4	7.2	17.7	15.8	49.5
Problem 2	32.4	16.2	29.7	13.6	18.6	22.7	47.2
Difference	-3.4	-3.6	1.7	-6.4	-.9	-6.9	2.3
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	.38	.27	.35	.23	.31	.24	.54
Problem 2	.33	.25	.42	.23	.27	.32	.33
Difference	.05	.02	-.07	.00	.04	-.08	.21

TABLE 19
SYNTHETIC LAG (EXPLICIT)

Problem	Absolute						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	-11.0	-4.6	-3.4	-2.6	-6.4	-24.0	-3.2
Problem 2	-30.0	-9.4	-4.2	-1.1	-17.0	-32.7	-42.7
Difference	19.0	4.8	.8	-2.5	11.5	7.8	39.5
Problem	Relative						
	All Groups	Group 1 (Ph.D.)	Group 2 (1st yr. College)	Group 3 (Nat. Sci. Ph.D.)	Group 4 (Other Ph.D.)	Group 5 (Nat. Sci. College)	Group 6 (Other College)
Problem 1	-.14	-.17	-.08	-.15	-.10	-.22	-.04
Problem 2	-.17	-.23	-.03	-.00	-.30	-.16	-.22
Difference	+.03	+.06	-.05	-.15	+.11	-.06	+.18

cated groups, and shorter for the technically more sophisticated groups than for the others. This appears to hold true for both versions of the index, but is less clear for the relative measure.

The correlation between this variable and Actual Redundancy, significant at the .01 level, is to be expected. Most of what occurs in this interval acts to increase redundancy.

The correlation between the two performances for this variable is .01 for the Absolute version and .09 for the Relative. The author is at a loss as to the explanation for this low correlation in view of the small group changes.

6. SYNTHETIC LAG (EXPLICIT)—the Absolute A-S Shift Point minus the Absolute NASI Point (Explicit). This index is stated both as (a) Absolute and (b) Relative. (Data for these variables are presented in Table 19.)

Correlations. Absolute Synthetic Lag (Explicit) correlates with Actual Redundancy $-.41$.

Discussion. This index is a measure of the failure of the subject to shift to synthesis as soon as the explicit yield from analysis is sufficient for solution of the problem to occur. The following considerations are relevant to an evaluation of this index:

1. If the A-S shift occurs before the NASI Point (Explicit) is achieved, the synthetic lag is negative. Such a premature shift may be adaptive if the subject does not find it subsequently necessary to return to analysis for further items of information or for the NASI (Explicit) before achieving solution of the prob-

lem, since one then concludes that the NASI (Implicit) was the basis of operation. If the subject finds it necessary to return to analysis (inversion) after the shift occurs, the premature shift appears to be inappropriate, indicating overly flexible performance.

2. If the A-S Shift occurs after the NASI is explicitly achieved, the Synthetic Lag is positive, and its magnitude is an indication of the extent to which further inquiry is required by the subject before realization that the information already acquired constitutes an adequate basis for solution of the problem.

3. Whether the Synthetic Lag is positive or negative, its absolute magnitude is a measure of the extent to which shift from analysis to synthesis is inappropriate, provided that the NASI (Explicit) appears in the performance.

For all groups, the synthetic lag is negative on both problems. Very seldom does one observe a performance in which it is positive. The extent to which the shift is premature does not appear to depend on educational level or specialized technical training or interests. The change in the magnitude of the lag which ensues as the level of difficulty of the problem is increased also appears to be independent of these factors. In general, the size of the lag increases as difficulty increases; that is, shift tends to occur more prematurely, except for group 3.

Most of the above discussion also applies to the relative version of the index. Note, however, that as difficulty increases the lag for group 3 becomes relatively smaller, while that for group 4 increases. Similarly, the lag for group 5 decreases while that for group 6 increases. Since educational level is relatively constant in each of these two comparisons, perhaps the greater technical sophistication of group 3 and 5 explains the fact that groups 4 and 6, in contrast to 3

and 5, shift relatively even more prematurely on the difficult problem. The correlation of $-.41$ between the Absolute Synthetic Lag and Actual Redundancy, which is significant at the .01 level, means that as the lag becomes more negative, redundancy becomes larger.

Probably a small negative lag would be associated with a less redundant performance than a small positive lag, since a positive lag precludes good guessing. One would also expect large lags, whether negative or positive, to be indicative of higher redundancy. These statements, of course, are speculative, since we have so little data on positive lag. The correlation between the two performances is .49 for the Absolute index and .40 for the Relative index. The tendency to shift to synthesis before analysis is completed appears to be fairly consistent for a given individual. Both correlations are significant at the .01 level.

H. Mean Values and Standard Deviations for All Variables

The mean values and standard deviations obtained on the two problems by

our full population are summarized in Table 20. It will be noted that not all the variables just defined are presented. Some of these variables are used in the calculation of others and for no other purpose. The correlation between performance on the two problems is also presented for each variable. Table 20 is then a summary of the data which has been presented under the heading "All Groups."

Once more let us emphasize that this set of data should not be regarded as an accurate indication of the consistency of PSI performance for the individual. Some aspects of PSI behavior are markedly changed when problem difficulty increases, perhaps due to the necessity for reorganization of process when the severity of the demands imposed by the

TABLE 20
MEANS, STANDARD DEVIATIONS, AND "RELIABILITY" OF VARIABLES
($N=45$, $r=.297$ for $p=.05$, $r=.380$ for $p=.01$)

Variable	Problem 1		r_{12}	Problem 2	
	Mean	SD		Mean	SD
<i>Work Variables</i>					
Time	23.8	16.8	.462	44.5	21.0
Questions	69.1	53.4	.462	110.0	86.8
Complexity	1.71	.23	.112	2.30	.69
Rate	3.01	1.25	.844	2.27	1.15
Pauses	.09	.16	.758	.18	.20
Changes of Set	1.5	1.2	.506	2.5	1.5
Percentage of Nonlinearity	12.6	11.8	.022	22.5	18.2
Effort	379.2	521.5	.424	814.5	733.5
<i>Information Variables</i>					
Exhaustiveness of Inquiry	13.6	2.5	.184	13.0	2.5
Actual Redundancy	.79	.10	.597	.86	.09
Inferential Lag (Absolute)	40.0	45.7	.297	62.7	74.9
Inferential Lag (Relative)	.52	.26	-.121	.50	.32
<i>Approach Variables</i>					
Analytic-Synthetic Shift (Abs.)	46.3	39.9	.126	47.9	35.0
Analytic-Synthetic Shift (Rel.)	.70	.21	.334	.55	.25
Synthetic Lag, Explicit (Abs.)	-11.0	35.8	.487	-30.0	68.1
Synthetic Lag, Explicit (Rel.)	-.14	.26	.398	-.17	.38
Synthetic Lag, Implicit (Abs.)	29.0	32.8	.003	32.4	34.1
Synthetic Lag, Implicit (Rel.)	.38	.27	.086	.33	.25
Predominant Mode (A/S)	6.32	7.58	.696	3.29	7.77
Frequency of Change of Approach	3.6	2.9	.042	7.8	5.8
Mixture of Modes (A minus S before shift)	.79	.20	-.387	.66	.24

task on the subject increases. To determine reliability, it will be necessary to construct two problems of more comparable structure than the two here used, and then to analyze the results

I. Summary of Discussion of Variables

Now that we have presented the full set of variables so far developed for the PSI, we may gain an overview of the analysis which they enable. In Fig. 5, we

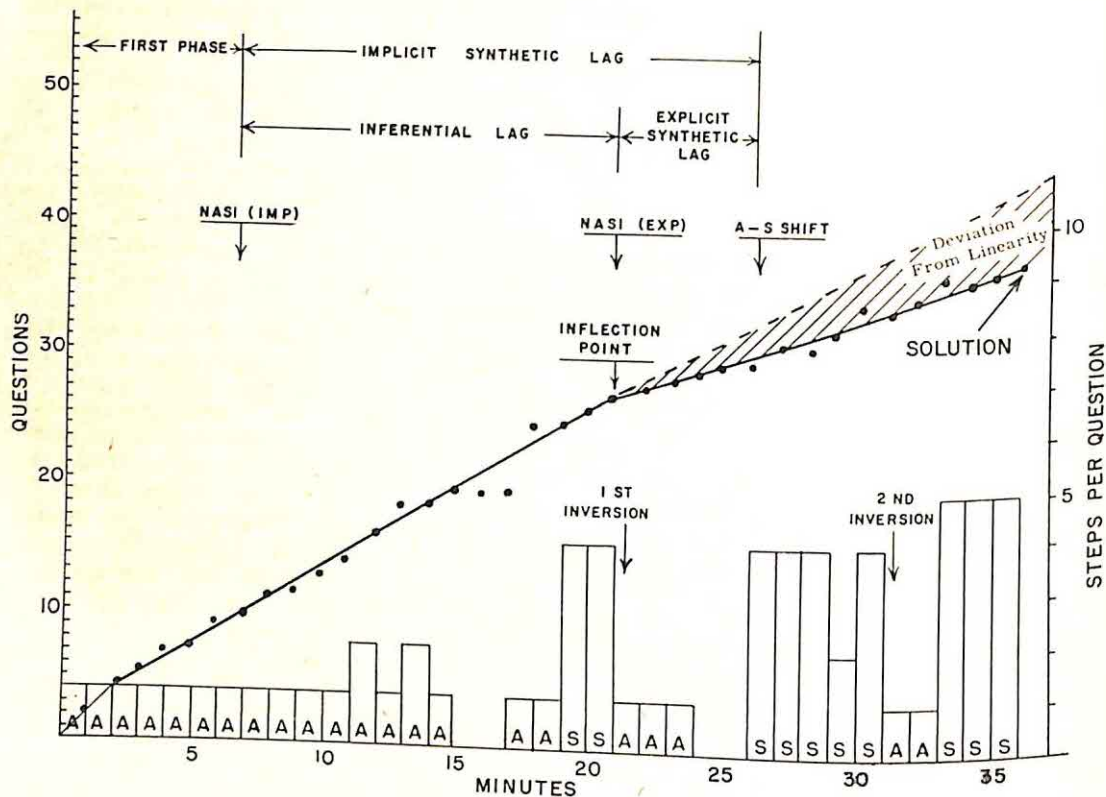


FIG. 5: Output and Analysis-Synthesis Graph. This illustrates the salient features on which PSI scoring is based.

obtained from the administration of these. While it is important to ascertain such data, this was not possible in this exploratory work. This shortcoming is to be regretted. We gain here, in addition to an insight into the effects of increasing difficulty, only an approximate idea of the consistency of an individual's behavior on repeated exposure to the PSI. In view of the known differences between the two problems, we might assume that reliability figures properly obtained should be at least as high as the correlations so far presented.

present a combined Output and Analysis-Synthesis Graph, on which a number of the variables defined and discussed in the previous pages are illustrated. Variables from all three areas of PSI performance are represented in this illustration.

The set of variables which we have defined gives information about a number of different aspects of the problem-solving process. Some of these variables appear to be sensitive to personality as well as to "cognitive" factors manifested during the performance. Such variables as inferential lag, synthetic lag, frequency of inversion, and even actual redundancy, to name a few, one would expect to be particularly susceptible to such factors as the effects of test

anxiety. A thorough analysis of the extent to which the variables listed above, which demonstrably describe characteristic and important parts of a person's problem-solving behavior, at least on the PSI and perhaps in other tasks, can serve as indices of personality organization is beyond the scope of this paper. Some definitely preliminary results relevant to these considerations will be mentioned in Section IV following, together with a brief and also preliminary survey of the relationship between some aspects of PSI performance and other less abstract cognitive tasks. The entire population of this study has been recalled for intensive psychological evaluation by Mr. Sidney Blatt, of the Department of Psychology of the University of Chicago, working in conjunction with Dr. Morris Stein of the same department. These same workers are also currently engaged in research to investigate the relationship between creativity in research chemists and PSI performance. Results from these researches will eventually be forthcoming from their laboratory.

On the basis of the work already carried out, it appears clear that much more is involved in PSI performance than cognitive factors alone. Examination of the extent to which stress will change characteristic performance on some of these variables, the extent to which others are relatively invariant across many conditions, correlative studies between this and other behavioral measures—all must be used before one can specify the dependence of PSI behavior on perceptual, personality, and cognitive elements.⁶

There is an obvious inadequacy to the variables so far developed, if our goal is to understand the manner in which the problem-solving process is generated by the interaction of an individual with a problem. Initially, the subject

has a choice between a large number of possible moves, with relatively little basis for discriminating between these as to utility. After the first questions asked of the PSI, the information which is forthcoming restructures the probabilities of various questions being asked next, so that instead of approximately equal probabilities for all questions, we now have a more probable and a less probable subset. The next choice is made, in a logical performance, from the more probable subset, with a consequent alteration of probabilities for all moves in accordance with the information derived from the two moves so far completed.

Ideally, we would like to be able to state why the n th step follows the n -first step, and in what way it derives from and relates to the previous elements in the process. This requires that we attempt to relate the state of the information possessed by the subject to the next step performed by the subject. We would hope ultimately to be able to include in the description of state those parameters of personality and perception and cognition which will affect choices between steps of approximately equal probabilities based on information considerations alone. If this is our goal ultimately—to understand the dynamics of the process through which each step is generated—we must acquire both an adequate description of state and a knowledge of the function by which the transition probability to the next step is determined. We have here stated a problem which is recognized today as that of analyzing a stochastic process with nonindependent transition probabilities. Our goal is to learn the function which stipulates the manner in which transitions from element to element of the process take place. It is hoped that the PSI will lend itself to this sort of analysis. The present author feels, however, that some of the variables so far developed represent largely static abstractions rather than dynamic functions, and will require much improvement before we can approach the above-mentioned goal.

III. PSI BEHAVIOR

A. General Description

A general description of problem-solving behavior as demonstrated on the Problem-Solving and Information Apparatus (PSI) can now be presented. This qualitative sketch represents a "typical" performance, to which numerous exceptions can be found in the data.

The starting point chosen by the subject for his initial attempts to structure the PSI

⁶To save printing costs, a discussion of the effects on PSI performance of education per se, education in a particular discipline, and interest in a particular discipline before specialized training, as well as development of a performance profile, an evaluation of our set of variables as comparative indicators, definition and discussion of a utility index, examination of the consequences of increased difficulty on our variables, and a comparison of PSI performance of a small group of college students with achievement test scores, college records, and personality indices, have been deposited with the American Documentation Institute. Order Document No. 5359 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D.C., remitting in advance \$2.25 for 35 mm. microfilm or \$5.00 for photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

situation is usually at the "output" end of the problem, and consists of a rapid and generally exhaustive investigation of the immediate antecedents of the output. Very seldom, one encounters an individual whose initial procedure is to permute the "input" elements in a systematic fashion. The problems are designed to make this approach difficult and tedious, and it usually gives way to the more systematic output analysis described above.

As the antecedent conditions for the output are determined, the rate of asking questions diminishes sharply and a slower and steady rate is established and maintained. At this point, many subjects commence verbalization which continues until solution is achieved, and start to take notes about the events which they observe. Verbalization, when it occurs, is of great utility in aiding understanding of the basis for certain aspects of the performance. Pauses occur during which notes and the apparatus are scrutinized and "self-discussed" aloud. These pauses are frequently followed by a transient acceleration in rate which persists for a time proportional to the length of the pause. The rate of reaching decisions appears to be very stable and characteristic of the individual, and is manifested by a steady linear increase in the cumulative question curve as drawn on the Output Graph. Some system of notation is adopted early in the performance, which serves as a shorthand rather than as a conceptual aid.

One might represent the progress of an individual through the PSI problems as a series of reformulations of the problem, each time focusing on different aspects of the many relationships involved. After the immediate antecedents of the output are discovered, the problem is restructured, and becomes in effect the problem of achieving these antecedents. Since there are several antecedents, they constitute what are in a sense subproblems. These are solved separately at first, and the result of the first restructuring tends to be concentration on the task of learning how to achieve these antecedents by using the permitted input elements. However, at this phase, the various antecedents are considered separately, and their achievement tends to be viewed as separate problems. The PSI problems used in the experiments reported here are so built that simple combination of the conditions which will achieve the antecedents separately results in an interaction between these conditions which precludes the solution from being achieved.

Recognition that the interaction between the simple solutions is such as to introduce a new factor into the constraints previously recognized as relevant in the problem is the prelude to a revision and reformulation of the problem as

perceived by the individual. It is at this point that most subjects encounter their first substantial difficulty. First, one must analyze the interaction in order to determine the manner by which solution is prevented. This phase of the problem is often accompanied by long and amusingly emotional verbalized arguments which tend to be syllogistic. Ultimately, the subject discovers the source of the blockage, and this discovery, even in staid adults, is greeted with great pleasure and satisfaction.

The discovery of the aspect of the interaction between the separate solutions to the subproblems initiates the final phase of the process, in which the subject devises a way to block the interaction. This consists of analyzing relationships in order to learn how to bring about a result which will prevent another result from occurring. This is the most serious source of difficulty in completion of analysis.

When information is acquired by the subject which might be expected to change set—in that it affords the opportunity to achieve closure of one of the aspects of the problem—an inflection point is often observed in the Output Graph; that is, the rate at which questions are asked changes to a new and also steady rate. This phenomenon is particularly of interest when the achievement of the necessary and sufficient information (NASI) for solution of the problem occurs implicitly; that is, when the information from which solution can be inferred is in the possession of the subject. One frequently observes a change in rate at this point, even though the subject continues analysis until he achieves these items of information (the NASI) overtly. This indicates that the significance of the information which was on hand at the point where the inflection occurred was not effectively incorporated by the subject even though it did cause a change of set. Surprisingly, the overwhelming majority of our sample went ahead to obtain the NASI explicitly before they would attempt to produce a solution to the problem by a sustained combinatorial approach. One tends to conclude that inference once drawn is confirmed by direct inquiry instead of by an attempt to apply the results of inference, thus testing validity implicitly. The inferential lag may be an index of the amount of reassurance needed before an inferred conclusion is accepted and translated into action.

Although the initial investigation of the antecedent conditions of the output is usually systematic, there is generally not a sustained methodical tracing-back of these antecedents through the network to the input end. A particular relationship is focused upon and clarified, and then attention is frequently directed to another aspect of the problem which is unconnected with the relationship previously investi-

gated. There is a tendency for the completion of what might be termed "blocks" of analysis, which relate to the subproblems referred to earlier. These blocks involve groups of related or contingent relationships. However, such block analysis is frequently interrupted by one or two ventures into unrelated aspects of the problem. This occurs particularly during the analysis of inhibitory relationships, where the subject must learn how to do something in order to prevent something else from occurring which prohibits solution. As mentioned earlier, this is the greatest source of difficulty in the problems. (Perhaps the sallies into unrelated areas which occur so often during this phase of the problem are to be considered as some sort of random attempt to avoid continued failure here. Often the unrelated actions are such as to elicit phenomena perfectly well known to the subject, and give somewhat the impression that after encountering frustration in the analysis of a difficult relation he consoles himself briefly by demonstrating his ability to control some aspects of the situation.)

Repetition of particular questions and elicitation of the same information in a number of different ways occurs to an extent which is surprising in a university population. Occasionally one observes repeated patterns of inquiry which are long and complicated but which appear over and over during the process. The over-all efficiency of the population, in terms of maximum utilization of the information content of events observed, is remarkably low. Even though subjects have decided to minimize questions rather than time (the instructions, the reader may recall, request that both be minimized), the use of inference to avoid asking a question is minimal. Rather, inference is tested by asking the question. Subjects give the impression of being unwilling to test an inference by applying the results to an attempt to achieve solution, as though they somehow attached more importance to such solution attempts than to other questions. Willingness to test an inference by using it in a synthetic attempt might be considered an index of willingness to take risk.

Separate analytic and synthetic phases can be clearly distinguished in the great majority of PSI performances. Generally, the clearer this distinction, the more effective is the performance. The shift to synthesis usually occurs prematurely; there is a tendency to try to make-do with insufficient information. Subjects will come back to analysis when forced to do so by their inability to produce a solution. The shift to synthesis occurs after surprisingly few questions in many cases.

Although performance is predominantly analytic, the shift point is located early enough on a relative scale to make clear that the necessity

to reconcile simultaneously three constraints, which is the logical requirement imposed in the two problems used in this work, is exceedingly difficult for this population. The synthetic aspect of the task contributes almost as much to the difficulty as does the analytic component.

Frequently, subjects were asked to describe the way in which their manipulations of the PSI contributed to the achievement of the output. It was observed that when synthesis was carried to a successful conclusion without a complete analysis preceding it, the subject would have difficulty in repeating his solution (which was always required) and might have a solution which was not parsimonious, i.e., a solution which contained irrelevant manipulations.

In such cases, frequently the subject would attribute completely erroneous functions to some aspects of the solution combination. When inquiry was pursued further, the subject would buttress his rationale by citing the existence of relationships which he had observed repeatedly during the analysis. On a number of occasions, the evidence so adduced was diametrically opposed to reality. Thus one gains the impression of a perceptual factor in the PSI situation, which enables a subject to see one thing, conclude that he saw the opposite, and subsequently experience the repeated contradiction of his conclusion by observed phenomena without revising the way in which he has structured the situation.

B. Some Effects of Increasing Difficulty

As the problem becomes more difficult, by what is essentially the addition of a single feedback relationship, a number of changes occur in the characteristic process. The time required for solution to be achieved almost doubles and the number of questions asked becomes 1.6 times as great. The variety of these questions actually decreases, so that we have a tendency toward greater stereotypy of response. This increase in stereotypy is such that where an item of information will be elicited about 5.4 times in the average performance on the easier problem, it will be elicited about 8.4 times on the more difficult problem. Not only does redundancy increase, but something like "rigidity" increases also, as indicated by a tendency toward perseveration of useless activity on the more difficult problem. These changes occur in spite of the fact that subjects almost invariably take notes.

The complexity of the average question increases markedly, even though the relationships being analyzed are of the identical complexity in both problems. The average rate decreases appreciably, and the percentage of time spent with no manipulation of the PSI doubles. The number of changes of set goes up, and the amplitude of the effect of these changes increases.

Actual redundancy increases from 79% to 86%. The fact that an over-all increase of a factor of 0.6 in the total number of questions asked is accompanied by only 7% increase in redundancy is some indication of the extent to which analyzing the ordering of a process generates conclusions other than those obtained by more static methods of treating data. The absolute inferential lag increases by about 50%, but this lag represents about 50% of the total performance in both cases. In other words, about one-half of the performance is occupied by the elicitation of relationships which could be inferred on the basis of information already obtained, before the necessary and sufficient information (NASI) which enables a solution is acquired. Further superfluous questions may, of course, be asked after this point, and invariably are.

The absolute synthetic lag is negative in both problems. The average subject tends to shift to synthesis before he obtains the NASI (Explicit) but subsequently finds it necessary to revert to analysis to obtain this. This premature shift is greater on the more difficult problem. It is interesting that the shift to synthesis occurs in both problems at about the same point after the achievement of the NASI (Implicit), although of course this shift is relatively earlier on the more difficult problem. The similarity in the absolute locus of the shift appears almost as though shift occurs after a fixed and characteristic amount of analysis "whether the problem is ready or not."

The number of inversions from synthesis back to analysis doubles with the increase in difficulty, the performance becomes much less predominantly analytic, and the mixture of analysis and synthesis before the shift point is greater. All of these considerations support the conclusion that as the difficulty of the problem is increased, this group of subjects experienced relatively greater obstacles in that aspect of the task which required the synthesis of known relationships into an integrated product which met the definition of a solution.

Thus it appears that some aspects of the PSI performance are fairly constant and characteristic, while others are appreciably modified by increments in difficulty. The relatively invariant aspects of the behavior might reflect the intrinsic organization of the individual performance, and might give an insight into the extent to which factors related to personality influence the problem-solving

attempt. The more labile aspects of performance may be consequences of reorganizations of process under the impact of greater stress.

IV. APPLICATIONS OF THE PSI

A. Comparisons Between Groups

In order to clarify the factors which contribute to performance on the various Problem-Solving and Information Apparatus (PSI) variables, we have presented data collected from small groups of different composition with respect to educational level and area of specialized training or interest. These data have been used to estimate the dependence of our variables on these factors throughout the earlier sections of this paper. However, we may also use this material as a basis with which to evaluate the over-all contribution of education in general and of various kinds to PSI performance as a whole.⁶

A number of questions immediately come to mind when one considers the nature of the task involved in the PSI. Will education as such improve performance, and in what ways? Is training in the natural sciences more effective than training in other areas, and how is such increased effectiveness manifested? Can differences be demonstrated which are not dependent on such special skills of logical manipulation as are acquired in a university? What aspects of real-life performance are related to the behavior elicited by the PSI? Can one distinguish different modes of behavior which are concomitants of differences in personality structure?

The results which we have obtained clarify some of the questions asked above. We cannot claim to have provided definitive answers to many, nor did we expect to do so in as frankly exploratory a

study as this. Many of our comparisons are made with small samples which show consistent trends but do not reach statistical significance. Yet a clear picture begins to emerge from the data. Most of all, it becomes apparent that answers to these questions can be elicited using this technique, and areas which merit additional intensive inquiry become highlighted.

The over-all effects of education, as demonstrated by the results of a comparison between small groups of first- and eighth-year university students across our full set of variables, are to improve performance. Yet this improvement is so slight as to suggest that much more is involved than simply education per se. A comparison of Ph.D.-level individuals in the natural sciences with Ph.D.-level individuals in other areas of specialization shows that, on *each* index of performance, the natural-scientist group differs from the other group in a direction of greater effectiveness. This raises two possibilities: Either training in the natural sciences is the basis for superior performance of a group so trained, or perhaps individuals with a particular kind of habitual approach well-suited to this kind of problem are more likely to enter the natural sciences.

In an effort to ascertain whether the latter factor was operative, a group of college students was divided into two subgroups, one of which intends presently to enter the natural sciences for specialized study, and another which does not. Comparison of these two groups shows that in almost every respect, *the differences demonstrable between the two groups at the Ph.D.-level already exist between these two groups at the college level, before any specialized training.* It appears that an individual is likely to have a predictable approach and general behavior on the PSI based on his career interests. If one assumes that proficiency in the generic skills required in an area increases the probability that one will enter that area, it would appear that the activities involved in effective PSI performance are more related to the natural sciences than to other academic disciplines. The college-level group which intends to enter the natural sciences displays more effective behavior in a large number of variables than does the non-natural-scientist Ph.D.-level group.

If one compares groups with interests in the natural sciences at high and low educational levels, one can in effect partial out the contribution of differential interests to obtain an approxi-

mation of the effects of natural-science education on PSI performance, and similarly with groups in other areas of endeavor. The results of such an evaluation of the impact of different disciplines on PSI performance show that, while there is a definite quantitative increase in effectiveness after training, that increase is not as great as one would hope, does not span all variables measured, and is not accompanied by a qualitative change in the problem-solving process. That is, one concludes that for the present population, which is by no means representative, and for the PSI, which is a particular kind of problem, *advanced education serves to improve performance along pre-existing lines rather than to establish new and more effective problem-solving techniques*, and the amount of this improvement is not great.

The data raise the question of the developmental factors which contribute to the establishment of the differences which we observe before specialized training. A small number of experiments, which will not be reported in detail, carried out with extremely young (6- to 11-year-old) children, has shown that such technically naive subjects not only can solve PSI problems, but that the quality of some of their performances was better than that observed with some subjects at the Ph.D. level. It would be of great interest to determine whether two groups emerge in a population of children after a given level of experience or education is reached, whether such factors as home environment can be related to differential PSI performance, or whether perhaps two groups are found in the earliest measurements which can be made.

B. Relation Between PSI, Personality, and College Performance

The PSI would appear to require exercise of skills in analysis and synthesis. The results obtained from cross-disciplinary comparisons suggest that this is indeed the case, if one assumes that the natural sciences require skill in analysis and synthesis. In an attempt to obtain a more direct demonstration of the relation between facility with the PSI and facility in the performance of real-life tasks which require analysis and synthesis, we analyzed the performance of 16 college students on all comprehensive examinations taken during their residence at the University of Chicago. These examinations have subtests which measure the ability of the subject to elucidate and apply principles in the area of the examination. Examining the relationship between certain variables of PSI performance and scores on all such examination subtests in the subject's college career to date, we find a clear correspondence between effectiveness on the

PSI and the ability to do well on examination sections which require analysis and synthesis. Based on a trichotomy of high, medium, or low performance on the criterion of actual redundancy on the more difficult problem, a better ordering was obtained between relative college grades on the one hand, and cumulative analytic and synthetic subtest scores on the other, than by using either ACE scores or college grades alone.

Finally, preliminary examination of the relationship between differential PSI performance and some aspects of personality, carried out on the same sample of 16 cases by Mr. Sidney Blatt of the Psychology Department of the University of Chicago, shows a clear relationship between these factors.⁶

V. SUMMARY AND CONCLUSIONS

A technique has been developed and herein described and discussed which permits the detailed observation of the problem-solving process rather than its product. The technique involves the use of an apparatus which permits the presentation of many different abstract problems of quantifiable information content and structure. This apparatus is called the Problem-Solving and Information Apparatus (PSI).

A technique of constructing and scoring problems using the PSI has been described and discussed which enables objective measurement of the phases of acquisition, organization, and application of information by an individual during the process of achieving a solution to abstract problems. These aspects of the problem-solving process are related to the solution in a causal fashion, so that the solution is seen to be the outcome of a unified and cohesive sequential process.

Using 59 university students and staff members at different educational levels and with different areas of specialization, the dependence of variables in the areas of work habits, information acquisition and handling, and consistency and appropriateness of approach to the problem, on such factors as education, speci-

alized technical training, and career interest has been indicated.

To the extent that real-life facility in the solution of problems of the generic sort represented by the task which is posed by the PSI may be inferred from observation of PSI performance, a number of conclusions are warranted by the data. First, marked differences exist between groups of individuals engaged in the advanced study of certain disciplines. Second, these differences exist qualitatively between groups intending to enter these disciplines, before specialized study has occurred. Third, the effect of education varies according to the discipline studied, with performance in the various areas of the problem-solving process being affected differentially.

The relationship between PSI performance and ability to elucidate and apply principles in college courses has been demonstrated to be appreciable. It would appear that difficulty with such manipulations constitutes a major and severe constraint on the performance of all groups studied.

The data indicate that training or experience in certain activities, whether acquired early in life or during the course of academic studies, bring about habituation of an individual to certain kinds of conceptual and organizational processes which are consistently displayed in repeated PSI performance. These processes may be more or less appropriate to the requirements imposed by this sort of problem. In addition to such habitual processes, some aspects of personality appear to be reflected in the problem-solving process, such as, for example, self-confidence, anxiety, and compulsiveness. Personality factors as well as cognitive factors contribute to the PSI performance.

Further investigations of the problem-solving process using this technique are necessary before definitive analyses of the factors involved in performance can be attempted. Many potential applications must await the completion of such research. The technique may be useful for the development and evaluation of teaching and remedial techniques in certain fields, by evaluation of the performance of individuals before and after training in particular skills. By appropriate construction of parallel forms both containing and not containing specialized information, it might be possible to evaluate performance in certain professional and technical areas in a diagnostic way, so as to ascertain the extent to which difficulty in the manipulation of abstractions rather than deficiency in knowledge was the major constraint on the technical competence of an individual. The relationship of performance on this task to performance on standard psychological assessment devices must of course be determined, but perhaps of equal interest would be elucidation of the contribution of the kind of behavior here observed to abilities and skills used in less artificial situations. The method might be useful in evaluating the effectiveness of chemotherapy or other procedures on the ability of disturbed individuals to organize

cognitive processes.

In view of the clear differences which have been demonstrated between different groups entering college, and the relation of these differences to subsequent choice of career, it would be particularly interesting to gain an understanding of the factors which contribute to the establishment of these differences and the acquisition of these skills, the age level at which they are acquired, and the kinds of experience which are most conducive to their acquisition and effective retention. The results of such developmental studies might enable more appropriate preparation for those areas of intellectual endeavor where this sort of skill is of importance, by facilitating the construction of generalized methods for increasing effectiveness on tasks of this generic type. Present academic training does not appear to change parameters of effectiveness qualitatively to any appreciable extent.

Finally, the possibility of devising simpler tasks for laboratory animals, which permit the same sort of sequential observation and analysis of cognitive behavior, should be investigated. Present behavioral indices tend to be product measures, and possible data of value may be obscured by the failure to construct indices which give more insight into process itself.

BIBLIOGRAPHY

1. BLOOM, B. S., & BRODER, L. J. Problem-solving process of college students. *Suppl. educ. Monogr.*, No. 73, July 1950.
2. DUNCKER, K. On problem-solving. (Trans. by Lynne S. Lees.) *Psychol. Monogr.*, 1945, 58, No. 5 (Whole No. 270).
3. GLASER, R., DAMRIN, D. E., & GARDNER, F. M. The tab item: A technique for the measurement of proficiency in diagnostic problem solving tasks. *Educ. psychol. Measmt.*, 1954, 14, 283.
4. GRINGS, W. W. A methodological study of electronic trouble shooting skill. Rep. No. 9. Contract NONR 282-02, USC, LA, 1953.
5. HARLOW, H. F. *Symposium on psychology of learning basic to military training problems*. Rep. No. HR-HTD 201/1, Panel on Training and Training Devices, Comm. on Human Resources, Res. & Dev. Bd, May 1953, pp. 165-78.
6. JOHN, E. R., & RIMOLDI, H. J. A. Sequential observation of complex reasoning. *Amer. Psychologist*, 1955, 10, 470. (Abstract)
7. RIMOLDI, H. J. A. A technique for the study of problem-solving. *Educ. psychol. Measmt.*, 1955, 15, 450.

(Accepted for publication March 14, 1957)

Some Dimensions of Interpersonal Relations in Three-Man Airplane Crews

BENJAMIN FRUCHTER, ROBERT R. BLAKE, AND JANE SRYGLEY MOUTON

University of Texas¹

CONSTRUCTION OF THE CREW INTERACTION SCALES

Introduction

ONE of the interesting characteristics of a number of modern technological devices and systems is that their operation requires the team work of several people. This consideration leads to questions concerning the extent to which the interpersonal relations, which are part of the basis of organizational efficiency, can be measured and predicted.

The purposes of this study were: (a) to construct a scale for assessing the interpersonal relations of three-man, B-47, medium-bomber crews consisting of aircraft commander (AC), observer (O), and copilot (P); (b) to determine the basic content or factors represented in the items; and (c) to estimate the reliability and validity of the resulting homogeneous scales.

Rationale for Selecting Items

Several conceptual approaches to group phenomena were considered in selecting items to be included in the initial form of the instrument.

Theory and experiments in the field of group dynamics (5) which have led to the development of concepts such as "group standard" and "group

cohesiveness" were reviewed, and items suitable for measuring these dimensions in B-47 crews were developed. The approach by Homans (3)—which relates the group work system to its social system through variables such as liking and amount of interaction—also was examined, with items being written to measure these aspects of group behavior. Recent developments in group therapy (7) that emphasize group and individual relations between members and the leader, and also among members, both from a choice and perceptual standpoint, and that have resulted in experiments on the accuracy of interpersonal perceptions in small groups were examined; and items suggested by developments in this area also were included. Scales that had been developed for measuring the relationships among B-29 crew members were examined; some items from such scales were included because of their demonstrated value in measuring interpersonal relations. Finally, the sociometric literature was surveyed for additional items that had been found useful for evaluating interpersonal relations in a variety of work situations.

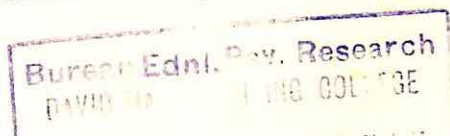
Items from each source were adapted for measuring relations among B-17 crew members. A final scale was developed after a tentative scale had been devised and evaluated through interviews with B-47 crew members and all items had been found acceptable in the sense that they assess relationships that crew members considered to be relevant in evaluating their crew performance. On the basis of nontechnical considerations concerning unsuitability of items for use in measuring crew relations, pointed out by flying and operational personnel, many items which otherwise appeared useful for scaling interpersonal relationships were deleted.

Description of the Crew Interaction Scale

The Crew Interaction Scale consisted of 44 items concerning a variety of aspects of crew relations.

The first 22 items required each crew member to rank himself and each of the other two members on some aspect of each member's crew

¹ This research was supported in part by the USAF under Contract Number AF 18(600)-602, monitored by the 3305th Research and Development Group (Combat Crew Training Research Laboratory), Human Resources Research Center, Randolph Air Force Base, Texas.



performance. Items included the extent to which each dominates the decision making of the crew, accepts responsibility for both air and ground activities, is agreeable and willing to go along with crew decisions, and exhibits self-control during periods of stress in the air.

The next five items dealt with how each crew member relates to the others. They included items such as the extent to which each crew member sees himself as clashing with the others, the degree to which each member feels he can influence the others to change decisions, and how much each feels a handicap in doing his job because of not obtaining necessary information and instructions from other members.

The final 17 items called for evaluation by its members of the crew as a whole. Items such as the degree to which the crew functions on the basis of an informal code of procedure and the extent to which members see the crew as a well organized team were included in this section.

The Ranking-Rating Technique

In a preliminary laboratory investigation with three-man groups, rankings had been found to be satisfactorily consistent from group to group (1). In addition to using the ranking method, it was also considered desirable to devise a rating scale for determining the *relative strength* of each ranking. The crew members were given the following directions:

In filling out the following items think of the three members of your crew—yourself and the two others—and try to characterize how each behaves in terms of the scale provided. Use the symbol AC for Aircraft Commander, P for Pilot, and O for Observer to identify yourself and the other two crew members. *Make an effort to put the identification symbol of only one crew member in any one box.*

Depending on the characteristics judged, two types of scales were used. One is a bipolar scale with a neutral point in the middle, and the other is an all-to-none scale with a 50-per-cent point at the center. In constructing a scale, the intervals were kept as uniform as possible by using the same or similar modifying phrases to define positions along the continuum. The following is an example of the type of scale used to

measure bipolar characteristics:

Rank the members of your crew for the extent to which they accept or reject responsibility in both air and ground activities.

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
always rejects responsibility	extremely often	very often
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
often	neither accepts nor rejects re- sponsibility	often accepts responsibility
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
very often	extremely often	always accepts responsibility

An example of a scale used to measure characteristics that can be scaled along an all-to-none continuum is:

Rank the members of your crew for the frequency with which they make errors in the performance of crew activities in the air.

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
always makes errors	extremely often	very often
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
often	makes errors some of the time	a little
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
very little	almost never	never makes errors

According to the instructions for both kinds of scales, each crew member placed his rankings of the three crew members (including himself) at the appropriate positions on the nine-point continuum. Tied ratings were not allowed. Some crew members objected to being forced occasionally to make a discrimination, as was required by the "no ties" restriction. The objection was met by allowing the rater to place more than one person within a scale interval and indicating the rank-ordering within the interval.

Items Included in the Scale

The following are the 44 items on which each crew member rated *himself* and the other two members of his crew.

1. Rank the members of your crew for the extent to which they accept or reject responsibility in both air and ground activities.

2. Rank the members of your crew in terms of being agreeable and willing to go along with crew decisions in both air and ground activities.

3. Rank the members of your crew in terms of the extent they take the initiative in getting things done in both air and ground activities.

4. Rank the members of your crew for the frequency with which they make errors in the performance of crew activities in the air.

5. Rank the members of your crew for the extent to which they work cooperatively with the crew in both air and ground activities.

6. Rank the members of your crew in terms of the amount of leadership they provide during air activities.

7. Rank the members of your crew for the extent to which they keep the crew "on the ball" in both air and ground activities.

8. Rank the members of your crew for the extent to which their actions are well-timed and appropriate to the actions of the other members of the crew in air activities.

9. Rank the members of your crew for the extent to which they place crew welfare above personal considerations in both air and ground activities.

10. To what extent do the members of your crew follow the directions and suggestions of the others in critical air situations?

11. Rank the members of your crew for the extent to which they put your crew at ease and make being a member of the crew more enjoyable in both air and ground activities.

12. Rank the members of your crew for the extent to which they feel frustrated by being a member of this crew.

13. To what extent do the members of your crew usually agree about what should be done in the air?

14. Rank the members of your crew for their competence, in air activities, in the specialty to which they are assigned.

15. Rank the members of your crew in terms of the extent to which they want to do a good job in both air and ground activities.

16. Rank the members of your crew for self-control during periods of stress in the air.

17. Rank the members of your crew for clarity in giving directions both in air and ground activities.

18. Rank the members of your crew for their satisfaction with the crew's over-all performance in both air and ground activities.

19. Rank the members of your crew for the extent to which they dominate the decision making of the crew in both air and ground activities.

20. Rank the members of your crew in terms of the extent to which they keep reminding the crew of regulations and SOP in both air and ground activities.

21. Rank the members of your crew in terms of the extent to which they are interested in their technical specialty in both air and ground activities.

22. Rank the members of your crew for the extent to which they are consistent (level of performance same from day to day) in their performance of activities in the air.

23. Rank the other members of your crew in terms of the degree you clash with them in crew discussions both in the air and on the ground.

24. Rank the two other members of your crew in terms of your liking for them in both air and ground relationships.

25. Rank the other two members of your crew in terms of the ease with which you can influence them to change their decisions in both air and ground activities.

26. Rank the other two members of your crew for the readiness with which you would expect them to back you up if a misunderstanding arose in your crew in ground activities.

27. Rank the other two crew members in terms of how often you are handicapped in doing your job because he has not given you necessary information or instructions in both ground and air activities.

28. To what extent does your crew work as a well organized team rather than as a collection of individuals in both air and ground activities?

29. To what extent do you feel "left out" of important decisions that your crew makes in both air and ground activities?

30. How well satisfied are you with the way you, personally, have worked out in this crew in its air activities?

31. What is your attitude toward remaining with your present crew?

32. Rate your crew for its enthusiasm for flying in present crew assignments.

33. Rate your crew for the speed with which it comes to agreement when decisions have to be made in the air.

34. Rate your crew in terms of its willingness to fly *under adverse conditions*.

35. Rate your crew for the extent to which it is a well organized team in both air and ground activities.

36. Rate your crew for its confidence in its ability to perform in critical situations in the air.

37. Rate the morale of your crew in both air and ground activities.

38. Rate your crew for the extent to which

it sticks together in the face of disrupting factors in both air and ground activities.

39. Rate your crew for the extent to which it is a "stickler" for following formal or official regulations and SOP in both air and ground activities.

40. Rate your crew for its effectiveness in planning together in both air and ground activities.

41. Rate your crew for the extent to which it is able to maintain a consistent level of flying efficiency.

42. Rate your crew for the frequency with which it makes errors in carrying out its air assignments.

43. Rate your crew for its "cockpit confusion."

44. Rate your crew for the extent to which it functions in terms of an informal code of procedures that it has worked out by itself in its air activities.

SAMPLE

The Crew Interaction Scale was administered to 90 male flying officers, who composed 30 intact B-47 crews in operational training. Crews had been organized for approximately four to five months. All were tested individually for an hour. Data collection took place over a one-week period since only small numbers of personnel could be made available in any one test period.

ITEM ANALYSIS

Agreement in Crew Ratings by Items

Since each member rated either the crew as a whole or specifically designated members, several ratings of a crew and its members were obtained from each item. From a statistical point of view the means of the several ratings should represent a better estimate of the characteristic being judged than any of the ratings considered separately. To represent useful estimates it is desirable that intercrew means be differentiated adequately relative to the individual ratings on the characteristic being judged. Horst (4) has developed a procedure which can be applied to estimating the extent of agree-

ment in ratings for any one crew relative to the discrimination between crews. If ratings are internally consistent, in the sense that the several judgments within each crew or by each crew member agree perfectly, the Horst coefficient would be $+1.0$. As the spread of the ratings around the mean of the within-crew ratings increases, the index approaches zero. Horst has demonstrated that the formula is related to the Spearman-Brown method for estimating reliability and to the Kuder-Richardson formula (no. 21) for internal consistency.

Each rating was weighted from zero at the "most desirable" end of the scale to eight at the "least desirable" end. Agreement indexes were computed for the following combined-score ratings. (a) The variance of the sum of the ratings of the crews on a given item was evaluated against the within-crew variance. (b) Next, the variance of the sum of the ratings of the crews on a given item was also evaluated against all the within-crew variance *exclusive of the self-ratings*. (c) Finally, the variance of the sum of the ratings for a *crew position* on a given item was evaluated against the within-crew variance of those ratings.

The Horst agreement coefficients for the scores included in scales are shown in Tables 5 through 10. The results for all of the items may be summarized as follows:

The interrater agreement indexes for the crew totals on the first 27 items (of which the first 22 were based on nine ratings and the next 5 on six ratings each per crew) were satisfactorily high with few exceptions. The results indicated both an acceptable consensus among crew members in rating one another and a satisfactory discrimination between crew means.

There were only three sets of ratings

for each crew on the last 17 items. As computed by the Horst method, these items had a wider range of interrater agreement values. Also, from the standpoint of the kind of judgments involved, items 29, 30, and 31 were different from the others, since the respondent was required to evaluate his own relationship to the crew rather than to rate the crew as a unit. Such items might be suitable for differentiating among individuals, but they would not be suitable for computing mean crew ratings. Indexes of agreement were therefore not obtained for these three items. Mean ratings for items 39 and 44 did not adequately discriminate among crews. Both refer to the extent to which the crew functions in accordance with a prescribed rather than an informal group structure. The agreement values for items 33, 34, and 40, concerned with speed in reaching crew agreement, readiness to fly under adverse conditions, and crew planning effectiveness, were also unsatisfactorily low.

The general conclusions from assessing the consistency of crew ratings by the Horst method were that crew members do agree in their judgments on most items and that there is satisfactory discrimination among the means of different crews. Items were next considered from the standpoint of validity.

Criteria Used for Validation

Two sets of proficiency rankings were procured from supervisory personnel at Lake Charles Air Force Base. They were a ranking of crews on a wing basis and a ranking of crews by squadron, within the wing.

The 67th Wing had been assembled as an operational unit four or five months previously. The Wing Standardization Board, which was composed of skilled personnel representing each of the three specialties, had completed the evaluation of the crews used in the validation

study during the week previous to the test administration. In making its evaluations the Wing Standardization Board used a fixed set of criteria, and, on the basis of the evaluation of specific components of individual performance and coordination among crew members, assigned a ranking for each crew in the wing.

Crews within each squadron were ranked in terms of performance efficiency as judged by three squadron commanders, who based their rankings on information supplied both by flight commanders and the squadron operations officers. The two most important considerations in making these ratings were the circular error in navigation and bombing assignments and the level of performance of the Observer.

Rankings by squadron and wing were used to distribute the crews into high, middle, and low performance groups to be used in validating the Crew Interaction Scale. The three performance groups were determined by combining the standings given by the Wing Standardization Board with squadron rankings. Each of the three sets of rankings was given an equal weight and the rankings summed into a single score. The 10 crews with the highest combined rankings were identified as the high group, the next 10 as the middle group, and the lowest 10 as the low performance group.

Item Validity

The validity for each item was determined in the following manner. An integral score value was located as close as possible to the median of the distribution for each item. The number of item-ratings above and below the median score for each of the three criterion groups was then determined. The following diagram shows how the data were set up for an item:

		Performance		
		Low	Middle	High
Median	Above	2	5	7
	Below	8	5	3

(Item 1 AC₁)

A chi-square test for the significance of the differences between the number above and below the median in the three criterion groups was made, and the contingency correlation coefficient, corrected for coarse grouping, was calculated for those items having significant chi squares.

The results indicated that many items are related significantly to the criterion groups at, or below, the 20-per-cent level of confidence. Item validity results will be discussed in connection with homogeneous scale validation in a later section.

CONSTRUCTION OF HOMOGENEOUS KEYS *Scaling Procedure*

To determine whether the items in the Crew Interaction Scale could be classified into several independent aspects of crew behavior, an analysis of the relationships among the items was made to identify homogeneous keys. The process was begun by classifying items into six categories, based on the rationale for test construction. Items were chosen for inclusion in the first tentative clusters on the basis of their agreement indexes, or significance levels of validity, or both. *The following set of symbols was used for designating scores:*

- a. The subscript *t* following the item number indicates the sum of all ratings made on the item (e.g. 1_t (The number "1" refers, of course, to item-number.)
- b. The subscript *t* following the abbreviation for a crew position for items 1 to 27 indicates the sum of the ratings received by that crew position (e.g., 1 AC_t).
- c. The subscript *s* following the abbreviation for a crew position for items 1 to 27 indicates the self-rating by crew position.
- d. The combination of the abbreviations of two crew positions separated by an oblique line indicates the rating(s) given by the one crew position to the other (e.g., 1 AC/O).
- e. On items 28 to 44 the crew positions' abbreviations following the item number indi-

cate the ratings given the crew by that crew position (e.g., 28 AC).

There were six tentative clusters identified at the beginning of the scaling procedure. They were composed of the scores listed below.

Cluster	Scores ²
1. Leadership	1 _t , 3 AC _t
2. Cooperation	2 _t , 5 _t
3. Interpersonal Relations	12 AC _t , 25 AC _t
4. Technical Competence	4 O _t , 15 O _t
5. Morale	13 _t , 33 _t , 36 _t , 38 _t
6. Formal Group Structure	21 _t , 25 P/AC, 40 P

After these preliminary scales had been assembled, scaling was done by Stice and Knoell's (6) modification of the Wherry-Gaylord (8) procedure for homogeneous keying. One advantage of this procedure is that it does not require the calculation of the item intercorrelations at the beginning of the analysis.

The correlations between the total score for each scale and score on each of the 86 items considered for inclusion in the scales were determined and evaluated. A correction was applied to the correlation obtained between the item and the scale score to compensate for spuriousness due to the inclusion of the item in the scale score. Items not in a scale, that had higher correlations with a scale than items that were included, were then added; and items having a low correlation with the scale were dropped. The process was continued in order to obtain a stable set of items for each scale. Iteration produced stable sets of items for four scales. Since the *Formal Group Structure* scale and *Interpersonal Relations* scale were not stable, no further analysis was made of them.

Approximately 26 item scores were included in the scales developed in the manner described above (see Tables 5, 6, 8, and 9). Additional homogeneous keys composed of items not already included in the existing scales were then derived. Item intercorrelations were estimated and placed in matrix form, and two additional tentative clusters were obtained by inspection. Items were added to clusters and clusters were stabilized through the iteration process. Two additional scales, identified by the titles *Crew Unity* and *Crew Coordination*, were derived (see Tables 7 and 10).

² See footnote to Table 1 for an explanation of the symbols.

TABLE 1
CORRELATIONS BETWEEN ITEM SCORES AND TOTAL SCORES ON THE SIX HOMOGENEOUS KEYS

Score ^a	Homogeneous Keys					
	Technical Competence	Leadership	Crew Coordination	Morale	Cooperation	Crew Unity
1 _t	.20	.51	.12	.45	.57	.50
1 AC _t	.07	.61	.17	.39	.36	.33
1 AC _s	.20	.31	.01	.10	.09	.20
1 O/P	.00	.48	.15	.16	.20	.33
2 _t	.21	.18	.07	.47	.67	.30
3 AC _t	-.00	.52	.28	.18	.17	.17
4 O	.62	-.11	-.17	.28	.23	.23
4 AC/O	.53	-.07	-.17	.22	.15	.13
5 _t	.22	.27	.07	.30	.75	.30
5 AC	.13	.20	.04	.33	.61	.17
5 O	.36	.20	.20	.30	.61	.37
6 AC/O	.36	-.18	-.17	.22	.09	.00
6 AC/P	.16	-.22	.01	.20	.20	-.07
7 _t	.13	.33	.28	.30	.25	.30
8 AC	-.07	.31	.48	.33	.04	.30
8 P/AC	.04	.19	.59	.40	.11	.17
8 P _s	-.14	.16	.48	.36	.14	.36
9 P/O	.13	.27	.28	.35	.44	.23
10 _t	.22	.10	.01	.37	.15	.27
10 O/P	.07	.09	-.12	.01	-.07	.03
11 AC _t	.18	.59	.29	.35	.44	.37
12 _t	.33	.27	.20	.69	.47	.40
12 AC _t	.49	.13	.07	.54	.47	.37
13 _t	.20	.18	.15	.35	.28	.23
14 O	.58	.07	-.01	.31	.28	.20
14 P/O	.36	.04	.12	.16	.20	.30
15 _t	.09	.33	.28	.39	.25	.62
15 O/P	-.09	.29	.23	.31	.25	.37
16 _t	.11	.38	.33	.54	.39	.50
16 P	.27	.22	.12	.56	.39	.53
17 _t	.29	.44	.28	.59	.39	.43
17 O/P	-.04	.29	.25	.12	.33	.07
18 _t	.56	.11	-.01	.37	.23	.07
18 P/O	.33	.27	.07	.26	.33	.20
19 _t	.13	.27	.28	.35	.44	.23
20 _t	.16	.27	.04	.33	.23	.33
20 AC _t	-.04	.36	.04	.24	.04	.17
20 AC _s	-.02	.31	-.07	.18	.15	.17
21 AC/O	.58	-.04	-.23	.18	.15	.13
21 AC/P	.44	-.04	.07	.47	.31	.23
22 _t	.24	.24	.33	.56	.41	.40
22 O	.47	-.07	-.01	.24	.12	.07
22 O _s	.40	.02	.12	.14	.09	.07
23 O	.27	-.07	.07	.03	.01	.00
23 P/O	.18	-.31	-.09	.09	-.15	.27
24 AC	-.11	.59	.39	.37	.33	.27
24 P/AC	-.16	.57	.33	.22	.12	.27
26 _t	.13	.16	.23	.49	.31	.40
26 O	.22	.18	.09	.41	.31	.68
27 _t	.00	.33	.17	.37	.64	.37
28 _t	.24	.07	.23	.55	.33	.30
29 _t	.18	-.31	-.09	-.09	-.15	.00
31 _t	.16	.20	.25	.45	.47	.50
32 _t	.22	.09	.15	.59	.39	.43
33 P	.11	.33	.31	.24	.15	.33
34 AC	.16	.00	-.04	.31	.04	.33
35 _t	.22	.36	.28	.70	.33	.40
36 P	.02	.18	.31	.20	.23	.33
37 _t	.36	.40	.25	.70	.39	.47
38 _t	.13	.24	.31	.47	.41	.27

TABLE 1 (Continued)

Score ^a	Homogeneous Keys					
	Technical Competence	Leadership	Crew Coordination	Morale	Cooperation	Crew Unity
38 O	-.04	.33	.45	.09	.23	.17
39 P	-.27	.33	.09	.09	.17	.17
40 _t	.13	.38	.20	.43	.36	.75
41 _t	.22	.33	.36	.63	.31	.40
42 _t	.34	.27	.31	.45	.41	.20
42 O	.07	-.02	.07	.07	.15	.17
43 O	-.29	.31	.56	.16	.07	.10

^a The following is the set of symbols for designating the scores in this table.

- The subscript *t* following the item number indicates the sum of all ratings made on the item (e.g., 1_t).
- The subscript *t* following the abbreviation for a crew position for items 1 to 27 indicates the sum of the ratings received by that crew position (e.g., 1 AC_t).
- The subscript *s* following the abbreviation for a crew position for items 1 to 27 indicates the self-rating by crew position.
- The combination of the abbreviations of two crew positions separated by an oblique line indicates the rating(s) given by the one crew position to the other (e.g., 1 AC/O).
- On items 28 to 44 the crew positions' abbreviations following the item number indicates the ratings given the crew by that crew position (e.g., 28 AC).

Item-scale correlations were examined and further changes were made in scale content by eliminating items having high correlations with more than one scale.

These operations resulted in the stabilization of scales identified as *Technical Competence*, *Leadership*, *Crew Coordination*, *Morale*, *Cooperation*, and *Crew Unity*. The loadings of the item total and part-scores on the six scales are shown in Table 1.

The item content of the scales, loadings, agreement indexes, and validity significance levels are given in a later section entitled "Interpretation and Discussion of Homogeneous Scales."

Intercorrelations of Scales

To determine the degree of interdependence among the scales, the six scale scores for each crew were intercorrelated by the product-moment method.³ The resulting correlations are shown in Table 2. An inspection of the coefficients indicates that the *Technical Competence*

³ The items were scored so that the end of the scale that was considered to represent the most favorable responses was given a weight of zero, and the response at the least favorable end of the nine-point scale was given a weight of eight. Therefore low scores were considered "good" scores.

TABLE 2
INTERCORRELATIONS OF THE SIX SCALE KEYS (MATRIX *R_s*)
(*N* = 30 B-47 CREWS)

Key	2	3	4	5	6
1. Technical Competence	.15	-.03	.57	.34	.52
2. Leadership		.62	.66	.68	.74
3. Crew Coordination			.63	.49	.68
4. Morale				.74	.96
5. Cooperation					.79
6. Crew Unity					

TABLE 3
MATRIX R_s^{-1}
(Inverse of Matrix R_s)

Key	1	2	3	4	5
1	2.4786	.3978	1.4564	-2.7437	.2077
2	.3978	2.4408	-.5329	-.8105	-.9336
3	1.4564	-.5329	2.7647	-2.4754	.3325
4	-2.7437	-.8105	-2.4754	5.8863	-1.6627
5	.2077	.9336	.3325	-1.6627	2.5881

scale has the lowest correlations with the other keys. *Crew Unity* was omitted from further consideration because of its high correlation with the *Morale* key. Of the two, the *Morale* key seemed more satisfactory. Scales were positively inter-correlated with the exception of a small negative coefficient between *Technical Competence* and *Crew Coordination*.

As a further test of independence among homogeneous scales, the inverse of their correlation matrix was computed. If the matrix could not be inverted it would be an indication that the scales were not linearly independent and some of the scales would be presumed to be combinations of others. The inverse (R_s^{-1}), as shown in Table 3, gave further support to the conclusion that the various scales (except "Crew Unity") represent essentially independent dimensions.

Scale Reliability

The reliability for each scale was determined by correlating odd and even item scores, based on the nine-point scale values. The Spearman-Brown correction for estimating the reliability of the full-length scales was also applied, and the corrected values are shown in Table 4. The reliabilities for the *Leadership*, *Crew Coordination*, and *Cooperation* keys are somewhat inflated since part scores from the same item appeared in both the odd and even total scores (e.g.,

two scores derived from Item 1 appear on the *Leadership* key).

With but one exception the corrected odd-even reliability coefficients are satisfactorily high for administrative use as separate scale measures. The exception is the *Leadership* key which, with a coefficient of .76, is satisfactory as a sub-scale within a battery.

Scale Validity

In order to assess the validity of the derived homogeneous keys, scale scores were correlated against the trichotomized performance criterion with results as shown in Table 4.

Since the means on the *Leadership* key for the middle and low groups were close together and since they both differed considerably from the mean for the high group, a biserial validity coefficient was also computed. The distributions for the two lower groups were combined and the resulting distribution contrasted with the distribution for the high group. A biserial coefficient was computed in the same manner for the *Crew Coordination* key. In both cases the biserial correlations obtained in this way were considerably higher than the corresponding triserials, indicating that the two scales differentiated the high group from the other two, rather than providing a uniform differentiation among the three criterion levels.

INTERPRETATION AND DISCUSSION OF HOMOGENEOUS SCALES

Technical Competence Key

Composition and definition. Five of the six scores composing the *Technical Competence* key shown in Table 5 represent ratings of the Observer by other members of the crew or by himself on certain of the items, as indicated. Three of the five scores represent total crew ratings of the Observer by all three crew members. The remaining two were ratings of the Observer by the Aircraft Commander. These ratings of the observer refer to aspects of his performance such as the frequency of errors, competence in performing the technical specialty, interest in performing the specialty, and consistency in performing the crew role. The crew's ratings of satisfaction with over-all crew performance (18 total) constitute the sixth valid item in the key.

Validity and interpretation. The key has a

positive, linear relationship of $+43$ with the criterion, with individual items ranging in significance from .10 to .20. The findings suggest that the greater the competence of the Observer in carrying out his assigned function, the more effective the crew, as ranked on the criterion. The finding that items that evaluate the competence of the Observer are valid is understandable, since a large portion of the criterion variance is considered to be due to the Observer's performance.

Leadership Key

Composition and definition. Five of the six scores that compose the *Leadership* key, shown in Table 6, are ratings of the aircraft commander. The sixth is a rating of the copilot. Items are ratings of the aircraft commander's ability to put crew members at ease, to make being a member of the crew enjoyable, and readiness to take initiative. The final item includes ratings of both the aircraft commander and copilot for the degree of acceptance or rejection of responsi-

TABLE 5
TECHNICAL COMPETENCE KEY

Item and Rating Used	Item Description	Loading	Interrater Agreement (Horst)	Contingency Coefficient	Significance Level of Validity (based on χ^2)
4 O (Total)	Rank the members of your crew for the frequency with which they make errors in the performance of crew activities in the air.	.61	.50	.50	.20
14 O (Total)	Rank the members of your crew for their competence, in air activities, in the specialty to which they are assigned.	.55	.62	.57	.10
21 AC of O (Total)	Rank the members of your crew in terms of the extent to which they are interested in their technical specialty in both air and ground activity.	.55	— ^a	.44	.20
18 (Total)	Rank the members of your crew for their satisfaction with the crew's over-all performance in both air and ground activities.	.50	.81	.50	.20
4 AC of O (See 4 above).		.50	—	.54	.10
22 O (Total)	Rank the members of your crew for the extent to which they are consistent (level of performance same from day to day) in their performance of activities in the air.	.48	.57	.45	.20

^a Dashes in the Agreement Index column indicate that it was inappropriate to compute this value since only one rating is involved in the score. Dashes in the Validity Level column indicate that the score was not valid at the .20 level of significance.

TABLE 6
LEADERSHIP KEY

Item and Rating Used	Item Description	Loading	Interrater Agreement (Horst)	Contingency Coefficient	Significance Level of Validity (based on χ^2)
1 AC (Total)	Rank the members of your crew for the extent to which they accept or reject responsibility in both air and ground activities.	.61	.42	.	.10
24 AC (Total)	Rank the two other members of your crew in terms of your liking for them in both air and ground relationships.	.59	.55	.7	.02
11 AC (Total)	Rank the members of your crew for the extent to which they put your crew at ease and make being a member of the crew more enjoyable in both air and ground activities.	.59	.35	.46	.20
24 P of AC	(See 24 AC above)	.57	—	.66	.02
3 AC (Total)	Rank the members of your crew in terms of the extent they take the initiative in getting things done in both air and ground activities.	.52	.51	.56	.10
1 O of P	(See 1 AC above)	.48	—	.64	.05

bility. The key seems to contain important aspects of the leadership function.

Validity and interpretation. All scores included in the present composition of the key have a significant relationship with the criterion. The key had a negative relationship with the criterion, and a triserial validity coefficient of $-.45$ was obtained for the total score. The mean for the high group was well differentiated from the means of the other two groups, and a biserial correlation of the high group versus a combination of the other two yielded a coefficient of $-.55$.

The negative relationship can be interpreted in several different ways. For the most effective crews, functional leadership may be based in the group as a whole rather than being exercised by the Aircraft Commander. Under these circumstances the Aircraft Commander would be less frequently required to put the crew at ease, to take initiative, or to accept responsibility. An alternative explanation of the negative correlation or inversion in mean ratings is that the members of the least effective crews felt threatened by a psychological test of the kind from which these data were drawn. They would feel compelled to depict their aircraft commanders as unusually competent. The least effective crews then might be less secure in admitting their own inadequacies so that they create a

"halo" of effectiveness for their aircraft commander by giving him more extreme ratings. The effect on ratings would be the same in either case. Perhaps both these types of causation occur.

A third possible interpretation is that those crews are most effective in which the aircraft commander does not put crew members at ease, does not try to make being a member of the crew more enjoyable, and does not readily accept responsibility. A similar finding has been reported by Halpin (2). He found that B-29 crews' ratings of the "consideration" of their aircraft commanders were negatively related to the effectiveness of these aircraft commanders in combat as rated by their superiors. He referred to this as "the dilemma of leadership," since the more the commander pleases his crew the lower he is rated in effectiveness by his superiors. In the case of the Crew Interaction Scale Leadership key, the crews that rated their aircraft commanders highest for making crew membership more enjoyable, accepting responsibility, etc., were least effective in terms of the performance criterion.

Crew Coordination Key

Composition and definition. The Crew Coordination key shown in Table 7 consists of ratings

TABLE 7
CREW COORDINATION KEY

Item and Rating Used	Item Description	Loading	Interrater Agreement (Horst)	Contingency Coefficient	Significance Level of Validity (based on χ^2)
8 P of AC	Rank the members of your crew for the extent to which their actions are well-timed and appropriate to the actions of the other members of the crew in air activities.	.59	—	.64	.05
43 O (Total)	Rank your crew for its "cockpit confusion."	.56	—	.57	.10
8 P*	(See 8 P of AC above).	.48	—	.65	.05
8 AC	(See 8 P of AC above).	.48	.11	.57	.10
38 O (Total)	Rank your crew for the extent to which it sticks together in the face of disrupting factors in both air and ground activities.	.45	—	.55	.10

of the extent to which the aircraft commander's and copilot's actions were well-timed and appropriate. Ratings by the observer of the crew's "cockpit confusion," and of the extent to which the crew sticks together in the face of disrupting factors, are included in the key. The common element among these items seems to be coordination within the crew.

Validity and interpretation. This key has an apparent curvilinear relationship with the criterion, the most effective crews having given themselves the highest, the least effective crews the next lower, and the middle crews the lowest mean ratings. A relationship with the trichotomized criterion of $-.23$ and a biserial correlation of $-.48$ were obtained.

The keyed items have validities that are significant at or beyond the 10-per-cent level, with curvilinearity apparent in the relationships of individual scores to the criterion. The curvilinear relationship might be explained by assuming that the most effective crews did not rate themselves highly coordinated because of their type of work organization, as discussed in connection with the Leadership key; whereas the least effective crews did not rate themselves highly because their coordination was inferior to that of the middle criterion group. On the other hand, since the means of the two lower groups are close together, the curvilinearity may be fortuitous, with the true relationship being one in which the upper criterion group is differentiated from the other two.

Morale Key

Composition and definition. All items in the Morale key shown in Table 8 are based on the combined ratings of the entire crew. Five items are crew-level items, while two others represent the sum of the individual ratings by the entire crew.

The items seem to fall into two subgroups. Ratings such as morale of crew, feeling frustrated, and enthusiasm for flying in crew assignments seem to be directly related to morale. Items such as maintaining a consistent level of efficiency, giving directions with clarity, and being a well organized team seem to be related to crew integration and to technical competence. Although items seem to refer to several aspects of crew organization, the common content of the items is designated as *Morale*.

Validity and interpretation. The total score for the key has a bimodal distribution and a curvilinear relationship with the performance criterion. None of the items has a significant degree of validity.

Cooperation Key

Composition and definition. The Cooperation key, shown in Table 9, refers to the extent to which members work as a cooperative unit, are ready to accept and adopt crew decisions, and feel affected because other crew members have not provided them necessary information or instruction. Three of the five scores, including

TABLE 11
SECOND-ORDER FACTOR LOADINGS OF THE FIVE KEYS

Key	Loadings				h^2
	Centroid		Rotated		
	I	II	I	II	
1. Technical Competence	.44	-.61	.04	.75	.57
2. Leadership	.77	.36	.84	.12	.72
3. Crew Coordination	.65	.41	.77	.01	.59
4. Morale	.92	-.22	.65	.68	.89
5. Cooperation	.83	.06	.73	.40	.69

and *Cooperation* keys have loadings on both of the second-order dimensions and seem to be complex at this level. While five homogeneous keys can be identified through item-scaling, results from the second-order analysis support the conclusion that there are two underlying factors operating to determine item responses. They deal primarily with judgments of the interpersonal working relations (*Crew Coordination* and *Leadership* keys) and with judgments of the technical competence of the Observer (*Technical Competence* key). A graphic representation of the relationships is shown in Fig. 1.

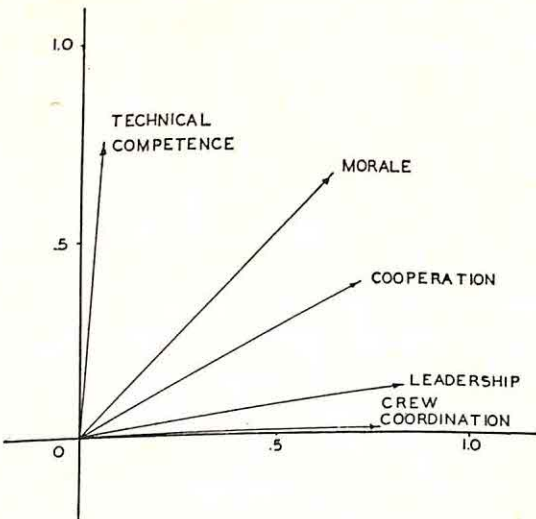


FIG. 1. Rotated positions of the five homogeneous keys on the two second-order factors.

Reliability and Validity of the Combined Scores

Since the relationships of some of the keys (*Leadership*, *Crew Coordination*, and *Technical Competence*) to the criterion, as well as to one another, are not linear, a combined score was obtained by an unweighted combination of scale scores, rather than by obtaining weights through multiple correlational analysis. The composite score, obtained by subtracting *Crew Coordination* and *Leadership* from the *Technical Competence* (Score = TC - CC - L), yielded a triserial validity coefficient of +.62, as shown in Table 12.

Since the high criterion group is discriminated from the other two better than they are from one another, a biserial correlation was also computed and yielded a validity coefficient of +.76. The corrected odd-even reliability of the Composite score was +.80.

Cooperation had loadings on both second-order factors, with the higher of its two loadings on the first factor. It also had a moderately negative validity coefficient. In order to determine whether the inclusion of the *Cooperation* key with its moderate validity would add to the validity of the composite, a second combination of keys was obtained by subtracting the *Cooperation* score from the composite score mentioned above (i.e., Score = TC - CC - L - C). It was anticipated that adding this key to the composite score would lower the validity of the composite, since the *Cooperation* key had loadings on factors with opposite relationships to

internal agreement was found for items 28 through 44, in which only three individual judgments were represented.

Item Validity

Based on a composite performance criterion composed from rankings of crews by the Wing Standardization Board and squadron commanders, item validities were computed to provide information concerning aspects of crew relations significant for the performance criterion. Approximately 95 scores with significance levels at or below .20 for the trichotomized criterion were identified.

Homogeneous Scale Development

Scales derived from the interrelationships of the ratings lead to the identification of five dimensions of crew relations which were identified by the titles *Technical Competence*, *Leadership*, *Morale*, *Crew Coordination*, and *Cooperation*. The scales representing these dimensions have satisfactory reliability as determined by the odd-even method and three of them show significant validity. One of the three validly predictive scales (*Technical Competence*) has a positive relationship to the criterion. The relationship of the other two validly predictive scales (*Leadership* and *Crew Coordination*) with the criterion is *negative*.

Second-Order Analysis of Scale Inter-correlations

To determine the manner in which the five homogeneous scales were grouped at the second-order level of complexity the scale intercorrelations were subjected to factor analysis. Two centroid factors were extracted from the scale intercorrelations. The analysis of the interrelationships indicated that there are two underlying factors on which ratings are based. One can be identified with the

Technical Competence key, which has a positive, linear, and valid relationship to the criterion. Except for further refinement, it seems to be satisfactory in its present form. The significant ratings on this scale were for the crew member called the Observer.

The other basic dimension concerns crew relationships. The *Leadership* and *Crew Coordination* keys have high loadings on it. For the *Leadership* key the high criterion group was differentiated sharply from the two lower groups. On the *Crew Coordination* key the tendency toward curvilinearity was even more marked, with the mean for the middle group being somewhat lower (representing more integration) than either extreme group.

Composite Scores

The composite scores obtained from the homogeneous keys of the *Crew Interaction Scale* take into account the opposite relationships of the basic dimensions with the criterion. A validity coefficient of .62 (triserial correlation) and .76 (biserial correlation) is obtained from the subtraction (without weighting) of *Leadership* and *Crew Coordination* from *Technical Competence*.

One other key, *Cooperation*, was considered for inclusion in a composite score since it had moderate negative validity. However, the *Cooperation* key had appreciable loading on both second-order factors, which have opposite relationships to the criterion. Combining it with the other scores raised the reliability but lowered the validity of the composite.

Conclusion

Within the framework of the present study the research has demonstrated that the direct assessment by crew members

of one another constitutes an effective means for investigating and measuring interpersonal relations in three-man crews, and provides a means for predicting the efficiency of crews that have already been organized.

An additional significant finding is that

three keys, *Leadership*, *Crew Coordination* and *Cooperation*, were related negatively with the criterion. Possible interpretations for these unexpected results were presented in connection with the discussion of each of the keys as given in the text.

REFERENCES

1. BLAKE, R. R., MOUTON, JANE S., & FRUCHTER, B. The consistency of interpersonal behavior judgments made on the basis of short-term interaction in three-man groups. *J. abnorm. soc. Psychol.*, 1954, **49**, 415-418.
2. HALPIN, A. W. The leadership behavior and combat performance of airplane commanders. *J. abnorm. soc. Psychol.*, 1954, **49**, 19-22.
3. HOMANS, G. C. *The human group*. New York: Harcourt Brace, 1950.
4. HORST, P. A generalized expression for the reliability of measures. *Psychometrika*, 1949, **14**, 21-31.
5. LEWIN, K. *Field Theory and Social Science*. New York: Harper, 1951.
6. STICE, G. F., & KNOELL, DOROTHY M. A simple mean-difference technique for obtaining scales. *USAF Hum. Resour. Res. Cent. Res. Bulletin*, 1953, No. 53-36, iii, 53-56.
7. TAGIURI, R., BLAKE, R. R., & BRUNER, J. S. Some determinants of the perception of positive and negative feelings in others. *J. abnorm. soc. Psychol.*, 1953, **48**, 585-592.
8. WHERRY, R. J., & GAYLORD, R. H. The concept of test and item reliability in relation to factor patterns. *Psychometrika*, 1943, **8**, 247-264.

(Accepted for publication March 16, 1957)

Psychological Monographs: General and Applied

Aggression and Age in Relation to Verbal Expression
in Nondirective Play Therapy¹

DELL LEBO

*Richmond Professional Institute of the
College of William and Mary*

AND ELAINE LEBO

Summer Hill School, Richmond, Virginia

PROBLEM

THE PRESENT study was an attempt to derive postulates from a theoretical consideration of aggression and age in relation to nondirective play therapy and to make predictions from these postulates. These predictions were then related to the types of responses that may be expected in nondirective play therapy from children of normal intelligence at various chronological age levels and of nonaggressive, intermediate, and aggressive personality.

It also seemed advisable to establish norms of play therapy data for possible use in future experiments.

It was postulated that children would manifest their aggression in their verbal behavior and, further, that aggression would be reduced in older children due to the process of socialization. From these postulates predictions were made regarding children's speech in a nondirective play therapy setting. The following hypotheses were formulated and examined.

¹The experimental foreground of this investigation was developed at Florida State University, Tallahassee; the theoretical background was reified at Carter Memorial Hospital, Indianapolis; the two segments were then blended at the School of Clinical and Applied Psychology of the Richmond Professional Institute, Richmond. The study completed, it was used by the senior author as a doctoral dissertation at Florida State University in 1956.

The writers gratefully acknowledge the assistance of: Ralph L. Witherspoon and the staff of the Nursery School and Eugene Boyce and the staff of the Demonstration School, who kindly made subjects available for more than two years; Anders Sweetland and Arnold H. Buss, the latter from Carter Memorial Hospital, both experienced therapists skilled in the evaluation of protocols, who served as judges; James Norton of the Statistical Laboratory, Purdue University, who generously shared his time and his knowledge to further this investigation; and V. J. Bieliauskas, director, School of Clinical & Applied Psychology, who thoughtfully arranged the senior author's teaching schedule to provide time for writing.

1. Not only are aggressive children more generally aggressive in their speech habits than other children, but also they are more bullying, assertive, bossy, and exclamatory in their speech. This hypothesis was investigated by examining the following types of verbal behavior: aggressive statements, threatened playroom rules, expression of definite decisions, and exclamations.

2. Since the imaginary dialogue of storytelling provides an excellent medium for aggressive expression and assertion, the speech of aggressive children contains more story units than that of other children.

3. Assuming that the aggressive child finds assertive experiences rewarding, it was hypothesized that in his speech he makes more favorable statements about himself, evidences more interest in the counselor, and makes more attempts to establish a relationship with the counselor than do nonaggressive children.

4. The relationship of the nonaggressive child to the counselor is presumably more one of dependency than of interest. Hence, it was hypothesized that nonaggressive children more frequently make unfavorable statements about themselves, express more indecision and doubt.

and make more attempts to shift responsibility to the counselor than do aggressive children.

5. It was also hypothesized that nonaggressive children have more conventional expressions of politeness in their speech than aggressive children. For example, aggressive children were expected to utter more "Hey!" and "Darn!" statements (hypothesis 1 above), while it was anticipated that the nonaggressive children would make more statements of "Hello," "Yes," and "Mmm," type.

In accordance with each of these hypotheses, based on amount of aggression, children with an intermediate amount of aggression should make statements whose frequency falls between the aggressive and nonaggressive children.

These statements made about aggressive and nonaggressive children were hypothesized upon the amount of aggression expected to be manifested in their speech during nondirective play therapy. Similar hypotheses were made about children at different age levels based upon the amount of aggression expected to be manifested at those levels. These hypotheses were as follows:

6. Children's speech is especially aggressive during their fourth and sixth years. Though children may be more aggressive during their fourth than sixth year, such aggressiveness does not generally find verbal expression as much during the fourth year as it does during the sixth, when speech habits are more developed. As the child continues to advance chronologically and, presumably, is subjected to increased socialization, the aggressive aspects of his speech are tempered. Consequently, at the ninth year verbal manifestations of aggression are still high, but not as high as they are during the sixth and fourth year of life. By the twelfth year the aggressive qualities of speech are negligible compared to the earlier years of life.

7. The younger children are expected to be more assertive or dominant than the 12-year-olds in their statements. In other words, younger children, because of their heightened aggressiveness, are hypothesized to make more of their own decisions about their playroom activities and attempt to shift fewer responsibilities to the therapist than would 12-year-old children.

8. Four-year-old children, because of their greater aggressive diffusion, test the limits of

the playroom more frequently than older children. The 12-year-old children, having been subjected to the most socialization, do the least amount of limit exploring. Twelve-year-old children confine their interests to routine questions indicative of curiosity or of a desire for information, rather than questioning how much they may get away with in the playroom. The difference in the verbalizations to be examined concerns statements of this type: It was hypothesized that the four-year-old child asks, "You can break the windows?" The 12-year-old, instead, asks such things as, "How come you brought me?" or, "Does this work like thus and so?"

9. The greater aggressiveness or dominance of children four, six, and nine years old results in attempts to relate to the counselor and evidence interest in him, rather than passively accept him. Presumably the older child waits for the counselor to make the initial moves in indicating the nature of the relationship. The 12-year-old, being less aggressive, is not so interested in establishing his position in the playroom. To the younger child the establishment of his place is important. Consequently, younger children are also expected to make more favorable statements about themselves than are older children.

10. Another way the more aggressive child can establish the assertive atmosphere he presumably finds rewarding is through the medium of stories. Hence, as a result of increasing socialization as well as more mature speech habits, it was hypothesized that children at the 12-year level tell fewer stories in the playroom than do children at younger ages. The six-year-olds are expected to tell the most stories since their speech habits are presumably superior to those of the four-year-olds.

11. The 12-year-old children are more polite as well as more adept conversationally than the younger children. Hence, when the 12-year-old speaks he utters more such phrases as, "Good-bye," and, "Excuse me," than younger children. Instead of telling aggressive stories the 12-year-old speaks of his family, school, pets, etc. to a greater extent than younger children. Similarly, it was hypothesized that the more socialized 12-year-old does less talking to himself and makes fewer sound effects as he plays than do younger children.

METHOD

Subjects for the present experiment were selected on the basis of chronological age, intelligence test score, and aggressiveness.

Chronological Age

To be considered chronologically representative a child could not vary more than four months from the age of 4, 6, 9, or 12 years. Since speech was the dependent variable of this investigation, children who could not play and talk at the same time were not suitable candidates for the experiment. The imperfect physical coordination of the three-year-old makes it difficult for him to perform two motor acts, such as dressing and talking, simultaneously. The four-year-old child can perform two such motor acts simultaneously. It has also been noted that at four years the child ceases the mere manipulation of toys and becomes interested in more constructive play activities. At this age also the child is said to become susceptible to flattery, to dread being made to feel ashamed, and to become concerned with the effect he produces on those around him. In brief, at about four years of age, the child's personality takes on a definite structure, contrasting rather sharply with the earlier undifferentiated picture. Consequently, it did not seem necessary to select children younger than four years of age for inclusion in the present experiment.

The lower limit of suitable age has been established by the maturational process. As for the upper limit of age, it was established on the basis of the general agreement that the possibility of successful play therapy decreases somewhere near the twelfth year of age.

Intelligence Quotient

To be considered intellectually satisfactory, a child's Stanford-Binet IQ score could not vary more than one standard deviation from the normal IQ of 100. This requirement was established in an attempt to circumvent the possible influence of extremely high and low intelligence on the types of statements children might make.

Beller Scales

Ratings of aggressive behavior were obtained through the use of the Beller scales (1). These scales are teachers' ratings of student classroom behavior. The scales include behavioral items on such factors as dependency, independence, dominance, submission, and aggression. They have been described as a reliable and significant measure, with "little overlapping among the rating scales used separately . . ." (1, p. 62).

The scale items determining aggression involve two items of verbal and two items of physical aggression. The highest positive correlation cited by Beller is that between these items. However, a study, such as the present one, concerned with types of statements rather than

physical indices of aggression should pay attention to the verbal habits of the children selected for study. More scale items concerned with verbalizations would seem to be a desideratum. Consequently, Beller's dominance scale, consisting entirely of items concerned with verbal behavior, was also employed. Beller's dominance and aggression scale items have been shown to have a statistical relationship. They also have a theoretical relationship which may be described in terms of the bossing and bullying of aggressive children. The aggressive child is also the verbally dominant child.

Additional evidence of the relationship between dominance and aggression is furnished by Griffiths (3). Griffiths' investigation seems to have equated characteristics called dominance by Beller with the total concept of aggression. Because of the emphasis on speech in the Beller dominance scales as well as the close relationship between aggression and dominance, both scales were employed in the present investigation.

Both scales have good reliability. Four pairs of raters achieved reliability figures ranging from .78 to .97 for items concerned with verbal aggression, and .81 to .97 on physical aggression. The scale items measuring dominance yielded r 's of agreement ranging from .86 to .97. Both scales include only behavioral items readily apparent from the child's interaction in the classroom. The scale items make no attempt to measure the child's attitudes toward his teacher or classmates.

In the present experiment the children's teachers served as their raters, and the children's classroom behavior furnished the source for their rated behavior.

Beller (1) obtained groups of children by selecting for his experiment those with high and low scores on the scales. A high score was defined as a mean rating score of 5 or more and a low score as a mean rating of 3.5 or less. The present experiment followed his method with the exception of introducing an intermediate group, i.e., those with scores falling beyond 3.5 but not reaching a score of 5. The intermediate group was selected in an effort to give continuity to whatever trends might be found.

Number of Subjects

Just as the senior writer did not rate the subjects of the present experiment, in order to avoid gaining any preconceptions about a child's aggressiveness, so he did not examine a child's Beller scales until the child's play therapy sessions were ended. Upon completing his play therapy experience the child's Beller scale was scored and he was assigned to the aggressive, intermediate, or nonaggressive groups.

There were 20, 22, 24, and 23 children aged

4, 6, 9, and 12 years respectively and 26, 27, and 36 children falling into aggressive, intermediate, and nonaggressive categories respectively.

Nondirective Play Therapy Sessions

Children found to be chronologically and intellectually suitable for inclusion in the present investigation were given three one-hour individual nondirective play therapy sessions with the same therapist in the same playroom. Three sessions were chosen because:

1. The child would have had an opportunity to become familiar with the playroom. His time would not be spent exploring the room.

2. The child would have selected his favorite play activities, if such a selection were to be made, and would be behaving in a characteristic manner.

3. The experimenter would no longer be a stranger to the child.

4. The child would have discovered the freedom of the room and would not play in a constricted manner if he did not wish to.

5. The meetings would not have become tiresome to the child.

The therapy records were kept in verbatim style, i.e., records were made during the play therapy sessions, both child and therapist responses were recorded, and the material was recorded exactly as spoken. This method was said to be the most reliable and complete when phonographic recording could not be used. It was expected that the inclusion of category W, to be described later, would make the record protocols even more complete than they would have been otherwise.

Borke Categories

The statements made by the children were categorized for comparison. The Borke categories (8) were employed for this purpose. These categories were originally developed by Helene Finke (2) and have twice undergone revision and expansion by the senior writer (7, 8). To avoid confusion, the categories employed in the present study have been given Helene Finke's married name, Borke. The differences between Finke's original, unpublished categories, the revised Finke categories, and the Borke categories are treated elsewhere (8).

One difference not specifically treated elsewhere is that of recording incomprehensibility in some children's speech. No previous investigation of the process of nondirective play therapy dealt with incomprehensible speech as part of the category system. Yet it seems unlikely that many investigators could have understood 100 per cent of what the children said. Studies ignoring incomprehensible speech should probably be

interpreted either as having recorded only comprehensible responses or as having ignored the incomprehensible remarks with such phrases as "words lost here." Category V of the Borke categories consists of sound effects, vocalizations which are not speech but which pertain to the child's play. Category W is deliberately set up to make mumbling or talking to self in a voice too low to be heard a part of the frequency count of the child's category usage.

Twenty-two of the 644 pages of verbatim style records were categorized by three experienced play therapists. These pages were selected on the basis of a table of random numbers. Finke (2), using five inexperienced judges, found that the percentages of agreement between her original analysis and that of the judges ranged from 66 to 77 per cent. In a previous investigation (7) the percentages of agreement between the criterion judgment and that of two experienced judges were 71 and 78 per cent. The percentages of agreement between the senior writer's judgment and that of two judges experienced in play therapy and skilled in the evaluation of protocol were 77 and 81 per cent for the present experiment.

In addition, it was possible to determine the similarity of category usage by an over-all measure of concordance among the three judges of the present study. The communality of judgments for the three observers was determined by the *W*-statistic (4). The coefficient of concordance of the rank order of each category was .84. This figure was found to be significant at better than the .01 level of confidence. It indicates a significantly high degree of similarity in the rank order of the judges' 662 categorizations. Thus, the categories have been found sensitive in the differentiation of statements made by children in play therapy.

The Borke categories employed in the present study are shown immediately below.

HELENE BORKE CATEGORIES FOR QUANTIFYING THE PLAY THERAPY PROCESS

A. *Curiosity about the situation and things present in it.* (Why did you choose me? Anyone else been here? Who owns these toys? Who drew that picture?)

B. *Simple description, information, and comments about play and playroom.* (This is an army. These are prisoners. More marbles. The room's different.)

C. *Statements indicating aggression.* (All references to fighting, shooting, storms, burying, drowning, death, hurting, destroying, etc.)

D. *Story units.* (1. Unconnected with play. Stories obviously farfetched or too exaggerated and inconsistent to have occurred. 2. Any

imaginary dialogue or story plot wound around the play, such as: He guards the opening. He's asleep. He doesn't know they're after him. I'm taking them to the army.)

E. *Definite decisions.* (I'm going to build a bridge. I said I'd do it and I did. Just what I wanted. Did it.)

F. *Inconsistencies, confusion, indecision, and doubt.* (My mother has two children, no, one. My brother is half my age and he's much taller. My sister's birthday was the day before mine last year but mine is before hers this year. I'm not sure what I should do. I wonder if this will work.)

G. *Exploring the limits of the playroom.* (Can I take this home? Can I get water? Can I paint this? I'm going to take this. One second. I can stay longer.)

H. *Attempting to shift responsibility to the therapist.* (What should I do next? Is this deep enough? Is this good? Do you like this?)

I. *Evidence of interest in the counselor.* (Were you here yesterday? What do you do? How are you? Can I trust you? Have you read such and such a book?)

J. *Attempting to establish a relationship with the counselor.* (Guess. Bet you can't guess. What's this? Look at that. See. Do you know what I'm going to do? Want to see how cars crash? Will you help me? You do this and I'll do that.)

K. *Negative statements about the self.* (I'm dumb. I'm afraid. I never win.)

L. *Positive statements about the self.* (I'm good in school. I can do that. I play marbles best. I'll win it back.)

M. *Negative statements about the family, school, things made or present in the playroom, the situation, activities, etc.* (Is there going to be new sand? I wish this was bigger. I don't like my sister. I wish I had more toys at home.)

N. *Positive statements about the family, school, things made or present in the playroom, the situation, activities, etc.* (I like it here. This doll is so pretty. We just got a wonderful new puppy at home.)

O. *Straight information and stories about the family, school, pets, teacher, self, etc.* (We have a big house. I went to the park yesterday. I have a sister. I was waiting for you. I thought you were my mother.)

P. *Asking for information.* (Do birds have ears? Where is the paint? How does this work?)

Q. *Questions or comments pertaining to time during the interview.* (How much longer do we have? I bet there are fifteen minutes left. Do I have time to play?)

R. *Exclamations.* (Here we go again! Hey! Darn! Oh! Crazy! Ahhh!)

S. *Unclassifiable.* (Yes. Mmmmm. OK. Hello. Goodbye. Excuse me. Any answer to a question

or a pure repetition of counselor's words.)

T. *Insightful statements revealing self-understanding.* (When I worried it made me steal. I wasn't loud but I was mean.)

U. *Ambivalent statements.* (I'm scared in here but I like to come here. I'd like to paint now and blow bubbles too.)

V. *Sound effects.* (Vocalizations which are not speech. Such noises as clucking, siren, machine gun, explosion, airplane, etc.)

W. *Mumbling or talking to self in a voice too low to be heard.* (Statements which cannot be heard and which the child does not direct to the therapist.)

RESULTS

The extensive data of the present investigation indicated that a normal distribution did not prevail for the categories.

Bartlett tests were applied to 16 of the categories. All but one of these tests were significant at better than the .01 level of confidence; the remaining test was significant at better than the .05 level. The data were transformed by both the square root and logarithmic methods in an effort to change the scale measurement to one in which the variance would be more homogeneous, i.e., less of a departure from normality. The square root transformation yielded significant Bartlett tests on nine of the categories. The log transformation gave 10 significant Bartlett tests. Six of the categories reached significance in both cases.

Evidently a statistical test that did not require the assumption of approximate normality was needed to test the hypotheses of the present experiment. The Kruskal-Wallis H test was used (5). This nonparametric statistic seemed to be the most sensitive of those available for application to the present data.

Results of a one-criterion variance analysis for ranks (based on frequency of Borke category usage) in relation to aggressiveness ($k = 5$) indicated that all of the categories save T and U reached significance. A similar analysis for rank in category usage in relation to age ($k = 4$) yielded significant H tests for all but four of the categories.

In every case of significance those children demarcated as *aggressive* had a *higher* rank-order (based on frequency of

TABLE 4

MEAN RANK OF PERCENTAGE OF CATEGORY USAGE BY CHILDREN IN FOUR AGE GROUPS,
WITH THE *H* TEST FIGURES AND THEIR SIGNIFICANCE LEVELS

Category	Age				<i>H</i>	Sig. Level
	4 (<i>N</i> = 20)	6 (<i>N</i> = 22)	9 (<i>N</i> = 24)	12 (<i>N</i> = 23)		
A. Curiosity.....	52	46	36	48	4.99	Nonsig.
B. Description.....	56	47	39	39	6.44	Nonsig.
C. Aggression.....	49	55	48	28	14.90	.01
D. Story Units.....	50	58	46	28	20.82	.01
E. Decisions.....	51	50	48	33	7.60	Nonsig.
F. Doubt.....	35	49	53	41	6.34	Nonsig.
G. Exploring Limits.....	63	50	40	29	23.79	.01
H. Shift Responsibility.....	46	39	46	49	2.01	Nonsig.
I. Interest.....	62	48	38	35	15.82	.01
J. Attempting Relations.....	64	50	39	29	22.55	.01
K. Negative About Self.....	44	56	47	34	9.12	.05
L. Positive About Self.....	54	45	40	34	8.70	.05
M. Negative About Others.....	44	52	44	40	2.58	Nonsig.
N. Positive About Others.....	44	49	47	40	1.76	Nonsig.
O. Information.....	48	51	40	42	2.71	Nonsig.
P. Asking for Information.....	52	38	38	54	7.71	Nonsig.
Q. Time Questions.....	60	54	35	34	17.40	.01
R. Exclamations.....	41	53	47	39	4.12	Nonsig.
S. Unclassifiable.....	23	39	49	66	30.25	.01
T. Insight.....	—	—	—	—	—	—
U. Ambivalence.....	47	46	49	38	4.16	Nonsig.
V. Sound Effects.....	52	57	43	30	16.00	.01
W. Talking to Self.....	43	55	52	30	17.23	.01

Note.—All mean ranks have been rounded to the nearest whole number.

gories, S (unclassifiable statements), reaches maximum relative usage by the nonaggressive children. Table 4 reveals 11 significant categories; of these, eight of the categories are significant also in Table 3. Category S, once again, is employed relatively more frequently by a nonaggressive group; this time it is the 12-year-olds.

Application of the *H* test to ranks based on relative frequencies would seem to sharpen its sensitivity to the extent of enabling it to pierce the massive verbal wall of the aggressive children and the 6- and 4-year olds. The results obtained from such usage will be those discussed in detail.

DISCUSSION

The outstanding finding of the present study would seem to be that aggression

and age exert a marked influence on the amount and variety of speech produced by normal children in the nondirective play therapy situation. So pervasive is this influence that future research in nondirective play therapy should seek to control these factors. Aggression and age should also be considered in theoretical formulations of nondirective play therapy. Rather clearly, not all subjects do make the same types of verbalizations in nondirective play therapy. The "regularity" of the process of nondirective play therapy is seriously questioned.

More specifically the results of the present study indicate that there is a difference in category usage of children noted as aggressive, intermediate in aggression, and nonaggressive as well as of children at different age levels.

On the basis of aggressiveness it was hypothesized that the children in the aggressive group would make more aggressive statements (category C) and threaten the rules of the playroom

(category G) to a greater extent than would children with teacher ratings of intermediate aggression or nonaggression. Table 3 indicates that both these differences are significant at the .01 level of confidence in the hypothesized direction. Aggressive children also voiced significantly more definite decisions (category E) and made more exclamatory statements (category R) than did children in lesser aggression categories. The first hypothesis received complete support by significant *H* tests.

Aggressive children also told significantly more stories (category D) than did other children. As was hypothesized, the aggressive children evidenced significantly more interest in the counselor (category I) and did make significantly more attempts to establish a relationship with the therapist (category J).

It is interesting to note the findings in regard to categories K and L (negative and positive statements about the self, respectively). The aggressive children made significantly more of *both* kinds of statements. This finding is interesting since it was hypothesized that it was the nonaggressive child who should make more K statements.

It seems possible that aggressiveness, in normal children, may be related to willingness for self-exploration, or at least willingness to verbalize opinion about the self. Consequently, in normal children aggression, like intelligence (6), may be related to insight. The important question of the nature of this relationship cannot be answered by the present investigation—first, because only normal children were used as subjects, second, because too few insightful statements (category T) were made to be considered.

In addition to the prediction that nonaggressive children would more frequently make unfavorable statements about themselves, hypothesis 4 also contained the expectation that aggressive children would make fewer attempts to shift responsibility to the therapist (category H) and would make significantly more expressions of indecision and doubt (category F). The differences in the usage of category H were nonsignificant (the intermediate group, in fact, had a slight increase over the aggressive group). Differences in expressions of indecisions and doubt were also nonsignificant.

It was further hypothesized that children designated as nonaggressive would make more conventionalized expressions (category S) than aggressive children. This hypothesis was supported by the *H* test.

Children with ratings of intermediate aggression and those in the aggressive group made significantly more comments about their play than did the nonaggressive children. As can be seen from Table 3, this was the only time a signifi-

cant difference was achieved with the intermediate aggression group having higher relative mean rank usage than the group designated as aggressive.

As was hypothesized, there was a significant difference in the employment of statements indicating aggression (category C) at various age levels, with the six-year-old children making the most such statements. In turn, as was also hypothesized, they were followed by the four-year-olds, by the nine-year-olds, and, lastly, by the 12-year-olds, who made the fewest such statements.

Twelve-year-old children were further hypothesized to make fewer definite decisions (category E) and to make more verbal attempts to shift responsibility to the therapist (category H). Neither of these hypotheses was supported by the *H* test (though the differences were in the expected direction).

Because the aggressiveness of the four-year-olds was said to be more diffuse, it was hypothesized that this age-group would make more statements indicating a testing of playroom limits (category G). Table 4 indicates that this hypothesis was borne out. With some slight departure from linearity a marked reduction is evident from one age-level to another. The 12-year-olds were also hypothesized to confine their curiosity to questions (categories A and P). The differences among the four age-groups were found to be nonsignificant for category A (curiosity about the situation and things present in it). The 12-year-olds, however, did employ this category more frequently than did the 6- and 9-year-olds. Category P (asking for information) was similarly nonsignificant. Here the 12-year-olds achieved a slight lead over all other age-groups in category frequency. Twelve-year-old children evidenced significantly less interest in the counselor and made fewer attempts to establish a relationship with him (categories I and J) than did children at other age levels, as was hypothesized. A significant difference in statements about the self was also obtained. As was hypothesized, the 12-year-old children made the fewest such statements.

Twelve-year-old children did employ significantly fewer story units (category D) than children at other age levels. The 6-year-old children made most use of story units, as hypothesized.

While 12-year-old children may be more adept at conversation than 4- and 6-year-old children, they did not offer more information about their family, school, pets, etc., than younger children. The *H* test figure for the category including such statements (O) was insignificant. However, on the positive side the 12-year-old children did utter significantly more polite but meaningless phrases (category S), did make significantly fewer

sound effects (V), and did less talking to themselves (W) than did children at younger age levels.

The significant group-differences in categories V and W are not surprising in view of the fact that their addition to the revised Finke categories was primarily to "serve to emphasize differences in speech habits between young and older children" (8). They were originally included on theoretical grounds. They received in this experiment their first empirical baptism. Interestingly enough the factor of aggressiveness did not influence them significantly.

Although no hypothesis was advanced concerning category Q (questions or comments pertaining to time during the interview) this category is one of the significant ones of Table 4. Perhaps category Q is significantly less frequently used by 12-year-olds for the same reason that S is employed more frequently. That is to say, perhaps the 12-year-old children realize it might be construed as impolite to ask, "How much longer do we have to stay in the playroom?" Usage of category Q may be a matter of aggression with the younger children who wish to assert themselves and to indicate a desire for termination by asking, "Is it time yet?" Or, perhaps, the answer is simpler than these theoretical speculations; the 9- and 12-year-olds may have had more wrist watches than the 4- and 6-year-olds and so may not have had to ask questions pertaining to time.

In summary, then, the information contained in Tables 3 and 4 indicates that:

1. The speech of the aggressive children contained significantly more comments about their play, aggressive statements, story units, definite decisions, explorations of the limits, evidences of interest in the counselor, attempts to establish a relationship with the counselor, negative statements about the self, positive statements about the self, and significantly more exclamations than did the nonaggressive group of children.
2. The speech of the aggressive children contained significantly fewer unclassifiable remarks (Yes, Hello, OK) than the speech of the non-aggressive group.
3. The speech of the younger children (those aged 4 or 6, and sometimes also those aged 9), contained significantly more aggressive statements, story units, explorations of the limits, evidences of interest in the counselor, attempts to establish a relationship with the counselor, negative statements about the self, positive statements about the self, comments pertaining to time, sound effects, and significantly more statements which could not be heard than did the children aged 12 years.
4. The speech of the younger children contained significantly fewer unclassifiable remarks (Yes, Hello, OK) than did that of the 12-year-olds.

It is interesting to notice that of the 11 significant differences indicated on the basis of aggression (Table 3), 8 of them also reached significance on the basis of age (Table 4). Once again the close correspondence between age and aggression is made manifest.

It would seem to be wise for nondirective play therapy to recognize and allow for the factors of aggression and age in its theoretical schema, if accurate predictions are to be made in that system. Without considering age and aggression the research methodology of nondirective play therapy may be seriously impeded. The recognition of the place of aggression in a child's normal development, both as a normal innate impulse and a normal expression of behavior, would be helpful in removing nondirective play therapy from the morass of procedural sameness that threatens it at present. The therapist seeks to modify the destructive element of aggression, while permitting the child to utilize this basic impulse in the service of a well-directed drive. It would seem that nondirective play therapy, to the extent that it sets realistic limits in an atmosphere of respect and confidence, already fulfills this therapeutic requirement. However, the theory of nondirective play therapy has been unaware of the role of the child's aggressive impulses and has failed to take into account differences in age. It is time both these factors were incorporated into its system.

SUMMARY

Nondirective play therapy, a method which arose from the postulates of client-centered therapy, does not formally consider the influence of aggression or age in its method. Studies of the process of nondirective play therapy, whether protocol-centered or child-centered, generally reveal the belief that children undergo

the same process, no matter what their age. The factor of age usually has been ignored or poorly controlled, while the factor of aggressiveness has not been considered.

From a theoretical consideration of aggression and age certain hypotheses were formulated. The hypotheses were then tested on a group of 89 children of normal intelligence selected on two bases. The first of these bases was that of age; there were 20, 22, 24, and 23 children aged 4, 6, 9, and 12 years respectively. The second basis for selection was teachers' ratings of classroom behavior. According to these ratings, 26 children were designated as aggressive, 27 as intermediate in aggression, and 36 as nonaggressive. These children were given three one-hour individual sessions of nondirective play therapy with the same therapist, in the same playroom. Verbatim style notes were made of their speech and vocalizations. These protocols were classified according to the Borke categories (8). It was according to these speech categories that specific predictions were made:

1. Aggressive children should make more aggressive statements, threats to playroom rules, expressions of decisions, and exclamations than nonaggressive children.
2. The speech of aggressive children should contain more story units than that of other children.
3. The aggressive child should make more favorable statements about himself, evidence more interest in the counselor, and make more attempts to establish a relationship with the counselor than nonaggressive children.
4. Nonaggressive children should more frequently than aggressive children make unfavorable statements about themselves, express more indecision and doubt, and make more attempts to shift responsibility to the counselor.
5. Nonaggressive children should have more conventional expression in their speech than aggressive children.
6. Six-year-old children should make the majority of aggressive verbalizations. They would be followed closely by those children 4 years old

who, in turn, would be followed by those 9 years old. The 12-year-old children should make the least use of speech falling into this category.

7. Younger children were hypothesized to make more of their own decisions about their playroom activity and attempt to shift fewer responsibilities to the therapist than the 12-year-old children.

8. Four-year-old children should test the limits of the playroom more frequently than older children. Twelve-year-old children should express more curiosity about the playroom and ask for more information than younger children.

9. Younger children should make more attempts to relate to the counselor, evidence more interest in him, and make more favorable statements about themselves than children 12 years of age.

10. Twelve-year-old children should employ fewer story units in their speech than younger children. Children 6 years old should use the most story units.

11. The 12-year-old child should utter more phrases such as, "Goodbye," and, "Excuse me," than younger children. The 12-year-old should also speak of his family, school, pets, etc., to a greater extent, and do less talking to himself and make fewer sound effects, than younger children.

The results of a one-criterion variance of analysis of the relation between category usage and aggressiveness, and between category usage and age, revealed significant differences in connection with a large majority of the categories. Aggression and age apparently made such an overwhelming difference in category usage that the aggressive children and the 4- and 6-year-old children employed more of all the speech categories demarcated as significant by the *H* test. These children, evidently, did so much talking that the specific hypotheses were obscured in a verbal fog. To discern meaningful differences, the *H* test was employed to study differences between ranks based on the *percentage-frequency* of use of each category by the three aggression-groups and the four age-groups, respectively.

From this analysis it was found that hypotheses 1, 2, 3, and 5 received statis-

tical substantiation in the three aggression-groups. Hypothesis 4 did not. The part of the hypothesis concerned with negative statements about the self not only failed to be supported but was flatly contradicted. The possible important applications of this contradiction were discussed.

From the analysis based on the percentage-frequency of use of each category by the four age-groups, it was found that hypotheses 6, 9, and 10 were supported. Hypotheses 7, 8, and 11 did not receive complete statistical confirmation. However, the trends were generally in the expected direction.

Most of the hypothesized differences in

speech categories were fully or partially substantiated. These substantiations suggest that the theories advanced concerning aggression and age can be utilized to predict the types of statements finding predominant expression in nondirective play therapy. The process of nondirective play therapy, as indicated by children's verbalizations in the playroom, does not seem to be the same for all children. Children of different aggression and age-groups respond differently, and in a predictable manner. Consequently, it is suggested that the factors of aggression and age require explicit consideration in the formal theoretical structure of nondirective play therapy.

REFERENCES

1. BELLER, E. K. Dependency and independence in young children. Unpublished doctor's dissertation, State Univer. of Iowa, 1948.
2. FINKE, HELENE. Changes in the expression of emotionalized attitudes in six cases of play therapy. Unpublished master's thesis, Univer. of Chicago, 1947.
3. GRIFFITHS, W. *Behavior difficulties of children as perceived and judged by parents, teachers, and children themselves*. Minneapolis: Univer. of Minnesota Press, 1952.
4. KENDALL, M. G. *Rank correlation methods*. London: Griffin, 1948.
5. KRUSKAL, W. H., & WALLIS, W. A. Use of ranks in one-criterion variance analysis. *J. Amer. stat. Ass.*, 1952, **47**, 583-621.
6. LANDISBERG, SELMA, & SNYDER, W. U. Nondirective play therapy. *J. clin. Psychol.*, 1946, **2**, 203-213.
7. LEBO, D. The relationship of response categories in play therapy to chronological age. *J. Child Psychiat.*, 1952, **2**, 330-336.
8. LEBO, D. Quantification of the nondirective play therapy process. *J. genet. Psychol.*, 1955, **86**, 375-378.

(Accepted for publication March 17, 1957)

Psychological Monographs: General and Applied

Individual Differences in Whole-Part Approach and Flexibility-Rigidity in Problem Solving¹RALPH H. GOLDNER²*University of Chicago*

INTRODUCTION

IN MOST studies of problem solving a general description of the problem-solving process has been arrived at in terms of the solution of the problem. There still remain several questions that would illuminate the problem-solving process in terms of the problem solver. For this investigation, there were selected two processes, related to "direction of problem solving" and to "adaptability." The first process, *Whole-Part Approach*, is concerned with the way the individual perceives the problem; the second process, *Flexibility-Rigidity*, emphasizes the way the person functions while solving problems.

MEANING AND DEFINITION OF
PROBLEM SOLVING

In this investigation, the meaning of problem solving is limited to one that would help explain the behavior of the problem solver. Brownell's (3) definition of problem solving was adopted. Accord-

ing to this definition, problem solving "(a) refers only to perceptual and conceptual tasks, (b) the nature of which the subject, by reason of original nature of previous learning, or of organization of the task, is able to understand, but (c) for which at the time he knows no direct means of satisfaction. Problem solving then becomes the process by which the subject extricates himself from his problem." In short, problem solving involves the goal-directed character of thinking, an obstacle to thinking, and a solution. But to delimit the definition further, this report restricts itself primarily and essentially to the *problem-solving process* in itself.

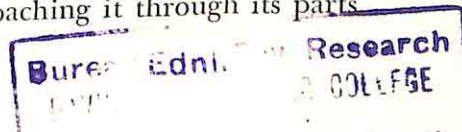
PROCESSES TO BE INVESTIGATED

Whole-Part Approach

The Whole-Part Approach determines one characteristic of the method used by the subject in problem solving. It has been noticed that some subjects approach problems as a whole while some others deal with parts of the problem. Thus, some persons, upon being shown a picture, will react to it in its entirety by calling it, e.g., a harvest scene, thus approaching the picture as an integrated unit or as a whole; others, seeing the same picture, will describe a segment or part of the picture by noting, e.g., a wagon, a man, corn stalks, thus approaching it through its parts.

¹ This monograph is based on a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy to the faculty of the University of Chicago. The writer wishes to express his deep gratitude to Benjamin S. Bloom for his patient direction and cooperation. In addition, appreciation is due Lee J. Cronbach, Ernest A. Haggard, and Ralph W. Tyler who assisted at various stages of the dissertation.

² Now at State University of New York Teachers College, Fredonia, New York.



Importance of the Whole-Part Approach

The Whole-Part Approach is an important problem-solving process for various reasons. It gives clues as to the individual's perception of the problem, his ability to organize parts into wholes. In addition, it seems to be related to personality variables and thus could account for individual differences in problem-solving behavior of people with similar intelligence-test scores.

Summarizing the discussion of earlier investigations which used concepts related to the Whole-Part Approach, the following findings seem to be important for the present study:

a. Several investigators have used the concept of Whole-Part Approach or similar concepts to describe the problem-solving behavior of normal and abnormal people.

b. There seem to be individual differences in the degree or extent to which normal individuals use the Whole or the Part Approach.

c. The Whole-Part Approach seems to be a process that can be observed in a variety of situations.

d. There is some indication that the Whole-Part Approach is consistent for the individual in a variety of situations.

e. It is a process that has meaning for everyday life situations.

f. It is a process that seems to have psychological meaning and seems to be related to personality characteristics.

Operational Definition of Whole and Part Responses

The review of investigations using the Whole-Part concept identified a variety of behaviors that have been taken as evidence of Whole-Part Approach in a variety of problem-solving situations. For convenience, the Appendix shows these behaviors classified as to Whole and Part Responses. This table will serve as a basis for arriving at an operational definition of Whole and Part Responses for the present investigation.

These earlier investigations seem to

suggest several aspects indicative of the Whole-Part Approach, namely, (a) manipulation of material, (b) speed of solution, (c) amount of material included in response or solution, and (d) plan of attack. The Whole-Part Approach as measured in this investigation combines these aspects. Considering the main aspects identified above, the characteristics of Whole Responses as well as Part Responses will now be described.

Whole Response. The subject usually does not manipulate material in a performance test but tries to think of a plan or a solution. He also proceeds to solve the problem quickly without further analysis. His responses include the whole area of the problem, or he may produce responses combining a great number of parts. Another characteristic is that the individual formulates a plan of action which he usually verbalizes.

Part Response. An individual using the Part Approach manipulates material readily and often produces responses in a random fashion. He usually takes a longer time working out a problem step by step. The subject responds only to one part of the problem at a time. He may break up the problems into parts and work out each one separately. He produces responses which combine only a small number of given parts. He usually has no particular plan of attack in mind, and when he verbalizes a plan it is only in terms of working out one part of the problem.

This report distinguishes between the words "Response" and "Approach" as used in Whole-Part thinking. A Whole Response or a Part Response is a single observation. Whole Approach or a Part Approach is the pattern resulting from a variety of responses. A Whole-Part Approach represents the over-all process that the person exhibits during problem solving. The conclusion is not to be drawn that subjects are classified as using either the Whole or the Part Approach. Rather, the Whole-Part Approach

represents a continuum ranging from preponderantly Whole Approach through Whole-Part Approach to predominantly Part Approach.

Flexibility-Rigidity Process

If one observes individuals solving problems in the course of their daily lives, one is aware of the great differences in the way people deal with problems. While some people will attack a problem in various ways, others are unable to shift to another attack even if their present performance does not bring the desired results. Some people can think only of one possible solution when they are confronted with a problem, while others are flexible enough to think of a variety of possible solutions. Certain personality characteristics and emotional states may limit the person's Flexibility, so that people with similar intelligence-test scores may be decidedly different in their problem-solving behavior.

Operational Definition of Flexibility-Rigidity

One of the purposes of this investigation was the identification of Flexibility-Rigidity as a problem-solving process. The emphasis was, therefore, on the observation of overt behavior which could be used to infer Flexibility or Rigidity.

Goldstein's (6) view of Rigidity seemed to apply mainly to impaired individuals showing an extreme degree of perseveration, while Fisher (4) limited the concept of Flexibility to situations where "alternative modes of behavior" are characteristics of a successful problem solution. In problem-solving situations both these aspects of Flexibility have to be considered. Not only must the subject continue to find another way of approaching the solution when his attack is unsuccessful, but at times he must find several possible solutions to a problem. Therefore,

the concept of Flexibility in problem-solving situations should include variety as well as variability of behavior. "*Variety*" would then stand for alternative solutions of the problems, while "*variability*" stands for behavior free from perseveration or behavior conducive to different attacks on the problem.

It is one of the hypotheses of this investigation that the structuring of a problem is an important aspect of the problem situation. If the problem is relatively unstructured as, e.g., in the Rorschach Test, variety and variability of response are possible and constitute an indication of flexibility. On the other hand, in several other tests, e.g., the Function Test, as mentioned later, the subject is *required* to arrive at alternative solutions and thus must show variety as well as variability of response in order to solve the problem.

In order to be able to deal with a number of different problems these aspects of flexibility have to be considered and must be included in an operational definition of Flexibility-Rigidity. For the purposes of this investigation it was found more advantageous to observe problem-solving behavior in terms of *Rigidity*, or *lack of Flexibility*. The operational definition of Rigidity used in this investigation is related to that of Werner (14), but has been adapted as follows: *Rigid behavior does not show variety or variability where variety or variability is either possible or required.*

It should be pointed out that the evaluation of variety of responses with respect to Flexibility-Rigidity has to be relative according to the particular population used in the experiment. Therefore, Rigidity of behavior does not necessarily mean pathological behavior but rather a decided difference of behavior from the rest of the subjects. Norms that have been set up to distinguish between normal and abnormal problem-solving behavior are not valid for this type of investigation that deals with the problem-solving behavior of normal subjects of superior intelligence.

Signs of Rigidity in the Experimental Situation

Before discussing the lack of Flexibility in the individual test situations, it is necessary to identify in general the behaviors that may be used to infer rigidity. Not all of the behaviors may be observed in the test situations. However, this dis-

cussion will serve as a guide to the observation of rigid problem-solving behavior during the experiment.

1. *Use of a similar attack on different problems*, as expressed through: (a) a decided preference for certain responses—in this study, an overemphasis on either the Whole or the Part Approach; or (b) an adherence to a particular attack on the problem—either (i) attending to the whole or part aspects of the problem by dealing first with the whole problem, then with the parts, and last with the minute details, or (ii) abandoning a progression from the whole to the part aspect of the problem and dealing with either whole or part of the problem exclusively.

2. *Limited production of responses*, in a problem situation in which the majority of subjects is able to produce many responses or solutions, is a sign of Rigidity. The amount of production that is labeled as rigid has to be determined on the basis of the performance of the majority of subjects.

3. *Absence or restriction of manipulation of problem material* in order to arrive at different responses is another evidence of rigid behavior, where such manipulation is appropriate or where it is generally true of other equal subjects.

4. *Lack of conceptual approach* is taken as a sign of Rigidity, especially in a person with more than average intellectual equipment. Besides changing the perceptual situation by manipulation of the problem material, the individual can change the conceptual situation by bringing to the problem other material that is not present perceptually or by looking at the problem in a different way. He may assume voluntary sets toward the problem; he may set up hypotheses which can be tested later on, or he may analyze the problem.

5. *Perseveration*. The degree to which a person can discard a response once made is indicative of his Flexibility in solving problems. There are two types of perseveration that have been observed in this investigation. In the first type the subject persists in giving the same response; he seems to be unable to restructure the situation, and he is relatively inactive in his problem solving. In the second type the subject may attempt some other responses but always comes back to the one he had made previously. Both types of perseverative behavior are indices of Rigidity and are treated together in this investigation.

6. *Lack of persistence in the exploration of the problem*. This is considered evidence of the inability of the person to restructure the problem.

7. *Presence of fluid, unorganized behavior*. This is characterized by the person's jumping from one aspect of the problem to another; by abandoning a product and starting again from the beginning; by verbalizing fluidity, as e.g., "My thoughts are so quick," "I forgot what I wanted to do," "After looking at it a while I get confused," "I am all lost."

These are the major signs of Rigidity that will serve as a basis for the evaluation of a person's Flexibility-Rigidity Process in the experiment.

PROCEDURE Population

The subjects were selected carefully with certain considerations in mind, namely high verbal ability, age, superior intelligence, and availability. It was believed that a selection of subjects who were homogeneous with regard to these variables would provide a critical test of the hypothesis that there are individual differences in problem-solving processes.

High verbal ability was important in this investigation in order to get sufficient verbalization during the experiments to infer problem-solving processes. Thus, it was decided to select a population that was older and that had a more than average amount of formal education. In addition, the population was to be of superior intelligence. Commonly accepted theories state that people with superior intelligence make greater use of the Whole Approach and also show a greater degree of flexibility than a less intelligent group.

It was finally decided to select first-year college students at the University of Chicago. By taking a first-year college group the chances were greater of getting a sufficiently large homogeneous sample with respect to the other factors mentioned above.

A population was selected with intelligence equivalent to the ninetieth to the ninety-fifth percentiles on the national norms on the American Council on Education Psychological Examination for College Freshmen (ACE).³ This test

³For the University of Chicago the mean derived score is 20 with a standard deviation of 4.

had been administered about six months before the experiments were conducted and it was assumed that no significant changes would have occurred in the six months' interval.⁴ It should be remarked that students with these scores represented a group slightly above the average of the general college population at the University of Chicago.

Another restriction to secure a more crucial test of the original hypothesis was that the Quantitative (Q) and Linguistic (L) score on the ACE Test should not differ significantly. It was thought that extreme differences in Quantitative and Linguistic scores may in some instances be an indication of either special abilities or of personal maladjustments, and might therefore introduce another variable into this investigation. Thus, only students whose Quantitative and Linguistic scores did not differ more than two points or one-half a standard deviation were selected.

To increase further the homogeneity of the sample, the results on two additional entrance examinations, that had been administered six months before the present experiments, were considered. These were the results on the Reading Comprehension Test and on the Writing Skill Test. It was decided to exclude students whose scores differed more than four points or one standard deviation from their respective total ACE scores. This would further eliminate people with special aptitudes, and provide for a more crucial test of the hypotheses.

The sample that was finally used in this investigation was selected from the records of the Board of Examination at the University of Chicago. Nineteen students, who fitted the criteria mentioned above, were selected to participate in the experiment: 12 men and 7 women.

Voluntary participation was believed to provide the proper motivation for this experiment. The investigator wrote a letter to each of the selected individuals inviting them to participate in the experiment at a time convenient for them.

The population selected had total scores ranging from 23 to 25. Therefore, the difference in total scores between 23 and 25 is only one-half a standard deviation.

⁴These records of the entrance examination were available to the investigator through the courtesy of the Board of Examinations at the University of Chicago.

This letter was followed up with a telephone call to assure participation of a large number of selected individuals.

Rationale for Choosing Tests and Their Description

In order to secure evidence for the inferring of problem-solving processes and to test the hypotheses about individual variation and individual consistency, the tests selected had to meet the following criteria.

1. Each test should involve a task requiring a maximum degree of overt action, including verbalization.

2. The tests should vary in degree of structure and should represent a variety of content. Variety in degree of structure is important in order to test the hypothesis that an individual is consistent with respect to his problem-solving processes in structured and unstructured situations. With a variety of task-content, greater and more accurate generalizations can be made with regard to the processes observed.

3. Each test should be unstructured to such a degree that individual ways of attacking the problem are possible—and thus, when they exist, individual differences can be inferred.

These are some of the major criteria to be considered in the selection of special tests for the observation of problem-solving processes. The following six tests were selected for the experiment because they appear to fulfill the criteria cited above. The tests are presented roughly in order of structure.

Rorschach Test

This test satisfied each of the criteria. It represents an unstructured situation in which the subject can attack the problem in a variety of ways and can give a variety of responses. He can react to the whole or only to part of the inkblot. The test contains perceptual as well as conceptual situations. The responses to the inkblot reveal the subject's perception of the blot. This is aided by the inquiry after the initial administration.

To what extent does this test represent a problem-solving situation? The subject has been asked to report everything he sees in the inkblot. He has not only to differentiate certain parts of the blot but he has to assign meanings to the percepts that are related or fairly close to the reality of the situation. He has to decide for himself how he wants to deal with the problem. No hints are given how the responses will be scored. It is, therefore, possible for the examiner to determine how the individual sees the situation, and how he deals with the problem.

Evidence of Whole-Part Approach. The evidence of Whole or Part Responses has been stated in earlier investigations. In this test the amount of material or area of the card included in the response is the determining factor:

1. *Whole Response:* the subject responds to the inkblot as a whole; he takes in the whole area of the blot.

2. *Part Response:* the subject responds to parts of the inkblot separately; he responds to minor parts of the inkblot.

It is the relative amount of Whole and Part Responses in all the 10 cards that determines the Whole-Part Approach of the subject.

Scoring procedure of Whole-Part Approach. All 19 records were scored according to Beck's method (2), and the number of W-, D-, and Dd-responses recorded. The number of W-, D-, and Dd-responses given by each subject were then divided by the total number of responses of the subject in order to determine the emphasis with which the subject responded to the Rorschach cards. Then the W%-scores of the group were arranged in a frequency distribution to determine the rating as to Whole-Part Approach of the subject. After considering Beck's (2) formulation of "normal" approach (20% W-67% D-13% Dd), and observing natural breaks in the distribution of W%-scores, ratings as to Whole Approach were given as shown in Table 1. By assigning these ratings, a person's Whole-Part Approach could be compared with his Whole-Part Approach in the other tests.

Evidence of Rigidity:

1. W%, 35% or more; D%, 80% or more; Dd%, 25% or more.
2. Number of cards with systematic sequence or with no succession: nine or more.
3. Number of cards with only one response.
4. No turning of cards when total responses are 35 or less.
5. Number of movement responses: three or less.
6. Organization score (Z): 20 or less.
7. F%: 55% or more.

TABLE 1
ASSIGNMENT OF WHOLE APPROACH RATINGS
IN THE RORSCHACH TEST

W%	Whole Approach Rating
45 and over	Very High (VH)
30-44	High (H)
15-29	Medium (M)
10-14	Low (L)
0-9	Very Low (VL)

8. Number of determinants used: seven or less.

9. A%: 45% or more.

Scoring procedure of Flexibility-Rigidity. Each record was scored according to Beck's (2) method. These scores were then arranged in a table and the nine categories giving evidence of rigid behavior examined. The decision as to rigid behavior was partly determined by a review of the available literature and also by the distribution of scores for this particular group, because the standards that applied to abnormal behavior were not always meaningful when applied to this group.

In order to arrive at a total score of rigidity for each subject, the number of categories showing rigid behavior were added for each individual. Thus, subject No. 6 showed rigidity in three out of the nine categories, and therefore received a score of 3. Although a weighting scheme was contemplated, an adequate basis for assigning weights does not seem to exist at this stage of our knowledge. The total scores were then arranged in a frequency distribution and ratings were assigned as discussed in a later section.

Function Test

Description. A test was needed that would represent a verbal conceptual situation, that would be relatively unstructured, and that would bring out a person's Whole-Part Approach. The writer, therefore, developed the "Function Test." This test consists of four problems. The task in this test consists of supplying several different uses for (a) a box, (b) a broom, (c) pliers, and (d) paper. It represents a relatively unstructured situation since no hints are given with respect to possible scoring, number of responses, or time limits.

Administration. The subject is asked: "What are the possible different uses of _____ (box, broom, pliers, paper)?" The investigator makes a verbatim record of the subject's remarks during the administration of the test. The subject deals with each object until he exhausts all the possibilities. There is no time limit. At the end of each problem the investigator makes a short inquiry to discover how the subject went about solving the task.

Evidence of the Whole-Part Approach. The analysis of the data indicates that an individual can deal with the object as a whole by leaving the object intact, or he can break up the object into parts and deal with the parts separately. Subject No. 7 responded to problem "Box" by dealing with it as a whole: (a) to stand on; (b) to fill it full; (c) mail packages; (d) waste-basket; (e) deliver in boxes; (f) mail box; (g) coffin; (h) cigarette boxes.

Subject No. 7 gave only one Part Response to "Box," namely "bonfires," thus breaking up the original configuration. Subject No. 11 gave as a Part Response to "Box": "Take it apart and make something else." His Part Responses to "Broom" were: "(a) take handle and use it as another implement—e.g., if you need another piece of wood; (b) straw to feed a horse or cow, or whatever eats it." Thus, in this test, those responses that described the use of the object as a whole are designated as Whole Responses, while responses that deal with the use of part of the object or that destroy the configuration of the object are called Part Responses. It is the relative amount of Whole and Part Responses in the four problem situations that determines the Whole-Part Approach of the subject.

Scoring procedure of the Whole-Part Approach. Each response by the subject was evaluated with respect to Whole Approach or Part Approach in accordance with the criteria stated above. The number of Whole and Part Responses was added, and a total score derived for each of the four tasks in this test. Then the total scores, which represented the number of responses, were computed for each subject with respect to Whole-Part Approach.

In order to arrive at an evaluation of the emphasis on the Whole or Part Approach for the individual, the number of Whole Responses was divided by the total number of responses and the percentage of Whole Responses computed. These percentages were then arranged in a frequency distribution and ratings were assigned to indicate the emphasis on W-Approach as shown in Table 2.

Evidence of Rigidity. This test is mainly a conceptual situation. In order to solve this problem the subject has to restructure the object and attempt to see its uses in a variety of situations.

TABLE 2
ASSIGNMENT OF WHOLE APPROACH RATINGS
IN THE FUNCTION TEST

W%	Whole Approach Rating
95-100	Very High (VH)
92- 95	High (H)
81- 91	Medium (M)
74- 80	Low (L)
Below 74	Very Low (VL)

A person who is flexible will be able to give many uses for the object, uses that are different in general idea (e.g., box: container, support, etc.). Thus, the number of general ideas offered becomes an index of flexibility. Subjects who offer only 12 general ideas or less in the whole test situation (four problems) are considered rigid.

Scoring procedure of Flexibility-Rigidity. Each response by the subject was evaluated with respect to the general idea represented. Thus, responses for the different possible uses of a box such as "to pack things," "for mailing," "as a nest," "as a tool box" were scored as belonging to the category "container." Responses such as "to sit on it," "as a foot rest," "to stand on" were scored under category "support." Among other categories were "fuel," "striking instrument," "wedge," "float," "writing surface."

The number of general-idea responses was added for each problem and the results for each problem examined in order to evaluate the individual's behavior with respect to rigidity. The scores for each problem were arranged in a frequency distribution, and ratings of High, Medium, or Low were assigned with respect to rigidity.

In determining these ratings, it appeared advantageous to assign the High and the Low ratings to the upper and lower 25% to 33% of the subjects, i.e., five to six persons at each end. It was thought important also to consider natural breaks in the frequency distribution, as well as the median score before ratings were assigned. A rating of "High" was assigned to the upper 25% to 33% of the scores, a rating of "Low" to the lower 25% to 33%, and a rating of "Medium" to the remaining scores in order to arrive at a Rigidity Rating for each problem.

In order to arrive at a total Rigidity score for the individual on the Function Test, the ratings of all four problems were added by assigning a value of 5 to a "High," 3 to a "Medium," and 1 to a "Low" Rigidity Rating. This method of totaling the ratings reflects the relative standing of the individual in the particular test situation.

The total Rigidity scores for the Function Test were then arranged in a frequency distribution in order to obtain a final Rigidity Rating for each individual on the Function Test. Again the upper 25% to 33% of the scores received a rating of "High," the lower 25% to 33% a rating of "Low," and the remaining scores a rating of "Medium."

Using the same method of scoring on the other tests, it was then possible to compare an individual's final Rigidity Rating on one test with his final Rigidity Ratings on the other tests.

Anagram I Test

Description. One way to observe process would be to put the subject in a situation where he would have an opportunity to manipulate materials. Sargent (13) had used anagrams. However, his subjects solved the problems without overt manipulation of letters. Therefore, the writer developed the "Anagram Test" consisting of two parts. The test consists of 14 wooden letters (C D E E I N O P R S S T U V), and the subject is asked to make words out of these letters. This was a relatively unstructured situation in which the subject could do what he thought was demanded by the situation.

Administration. The subject is told: "See what words you can make out of these letters!" There is no time limit but the test is stopped after about 10 minutes when most of the subjects have exhausted the possibilities. Any questions with regard to structure are answered with, "This is up to you." The investigator wrote down the words constructed. The observations included the amount of manipulation and placement of letters and any verbalizations.

Evidence of Whole-Part Approach. A subject using a Whole Approach in this situation will attempt to combine a great many letters in one word; he will produce big words. A person using Part Approach will be satisfied to produce smaller words. Thus, with respect to amount of material included in the response, the number of letters in a word is taken as a criterion of Whole or Part Response. In this experiment a word of five letters or more was considered a Whole Response. The division of five letters or more was arrived at arbitrarily after a study of

the test results revealed that the median was four letters. The words consisting of five letters and more represented 16% of the total responses given by all the subjects. Thus, at five letters there seemed to be a natural break in the distribution, distinguishing large from small words. A word of four letters or less was taken as a Part Response.

The Anagram I Test was selected as a less structured situation. Since the production in the first half of the test seemed to be less limited than that in the second half, the first half of the subject's responses was used to evaluate the person's Whole-Part Approach. The data showed a slight shift toward larger words in the second half of the test (see Table 3).

In order then to determine the person's Whole-Part Approach in a less structured situation, the relative amount of Whole and Part Responses in the first half of the individual's production has been used. An individual who starts out producing small words in the first half of the test may be forced to produce larger words in the second half because the production of small words becomes less possible.

Scoring procedure of Whole-Part Approach. In accordance with the criteria mentioned above, all the words containing five or more letters were scored as a Whole Response. A score for Whole Responses was computed for the first half of each individual's responses, dividing the number of words with five or more letters, occurring in the first half of the test, by the number of responses in the first half, as shown in Table 3. These scores were arranged in a frequency distribution and ratings were assigned with respect to Whole Approach. (Some individuals may remain at the same level as in the first part of the test, i.e., they keep on producing small words. Others produce small words in the second half after having produced bigger ones in the first half.)

TABLE 3
PERCENTAGE OF WORDS WITH SPECIFIED NUMBER
OF LETTERS: FIRST AND SECOND HALF
OF THE ANAGRAM I TEST

Part of Test	Number of Letters in Word					
	2	3	4	5	6	7
First Half (Responses = 467)	6%	41%	36%	12%	3%	1%
Second Half (Responses = 467)	4%	32%	40%	15%	7%	2%

Evidence of Rigidity. (a) Those persons who produce more than 50% of two- and three-letter words in the first half of the test are considered as showing a certain amount of constriction in the spontaneous reactions to the problem. (b) Those subjects who keep on producing 50% or more small words in the second half are considered as showing an even greater degree of constriction. (c) Those subjects who show a decided *increase* in production of small words in the second half of the test (over 10%) may have become more constricted when the situation became more structured; or they may have become less motivated.

The problem in this test is to make words. Nothing is said about the quantity, quality, or any other features of the words. Nothing is said about the manner in which the subject should proceed in constructing the words. He may continue to use the same letters and change the words mechanically (e.g., pit, pet, put, pot). If his production of words is restricted to this type of procedure he usually produces only small words. This behavior is taken as a sign of constriction in this investigation. If the percentage of words constructed by this perseverating method is 50% and above, it is considered evidence of rigidity.

Scoring procedure of Flexibility-Rigidity. The test was scored for each individual in accordance with the criteria stated above. The two- and three-letter words produced by the subject were added up for the first half of his production as well as for the second half. The number of small words in the first half was then divided by the total number of words produced in the first half and the percentage computed. The same procedure was used with small words produced in the second half.

Since the number of large and small words produced in the first half and in the second half was the same, the percentage of gain or loss in the production of small words was computed by comparing the percentage of small words produced in the first half with the percentage of small words produced in the second half.

Each record was scored with respect to the method with which the words had been produced:

1. Subject builds up words: e.g., pot—post; pin—spin.
2. Subject breaks down words: e.g., rinse—rise; pride—ride.
3. Subject changes vowel in the middle of the word: e.g., strive—strove; tone—tune.
4. Subject holds on to same letters and changes only the first letter: e.g., sip—nip—tip—dip; tries—dries—cries.

The number of words produced by each of

these four methods was then added and a score computed. In order to arrive at an evaluation of rigidity this score for each person was divided by the total number of words produced by the subject and the percentage of words produced by these four methods computed. The number of repetition of words was also noted for each person.

In order to arrive at an estimate of a person's rigidity in the Anagram I Test, the total scores of each category were arranged in a frequency distribution and ratings assigned, as discussed in the section above. Similarly, total Rigidity scores were computed and then arranged in a frequency distribution and final ratings assigned in order to make it possible to compare a person's behavior in the Anagram I Test with his behavior in the other tests.

Anagram II Test

Description. This situation represents a more structured one than the Anagram I Test. The subject is asked to construct one word out of all the 14 letters given (C D E E I N O P R S S T U V). These letters form the word "Productiveness." This word had been chosen because it was suitable for a variety of approaches and seemed to have the correct level of difficulty as determined by a few preliminary administrations. (Only one subject solved the problem during the actual experiment.) It also lent itself to the construction of several large words, an important aspect in the observation of the Whole-Part Approach.

Administration. The subject is told: "Try to make one word out of these letters!" There was no time limit but the test usually lasted 10 minutes. The subject was observed while he was trying to construct one word out of all the letters. During this time his moving of the wooden letters and his verbalizations were recorded. These records were then used to determine what the subject attended to and how he went about solving the problem.

Evidence of Whole-Part Approach. This represents a more structured situation, as the individual is asked to combine all the letters in one word. Besides considering the amount of material included in the response as in the Anagram I Test, the manipulation of the letters is taken as evidence of Whole or Part Responses. Lack of manipulation of the wooden letters and

"thinking of words" is taken as evidence of a Whole Response. On the other hand, ready manipulation of letters as well as construction of words in a random manner is taken as evidence of a Part Response. Furthermore, with respect to the amount of material included in the response, the production of prefixes and suffixes or working out a small part of the word separately is taken as evidence of a Part Response. Subject No. 10 used first a Part Approach by arranging letters as to vowels and consonants, and then by putting the two S's together. He used a Whole Approach by thinking of words like "vicissitude," "visualization." He verbalized this Whole Approach by saying that he was thinking of several words having sufficient length. After seven minutes of thinking of words he shifted to a Part Approach: "Maybe a trial-and-error process is better—and E in the back is probably . . .—assuming it began with a consonant an S as initial, for convenience let E follow it; let another consonant follow—N—; I have to put more than one consonant together—we'll run out of vowels!—'sent'—that doesn't recall a word."

The Whole-Part Approach of an individual was determined by the relative amount of Whole and Part Responses observed and recorded during the performance.

Scoring procedure of Whole-Part Approach. Each individual record was examined in accordance with the criteria set up above. Thus, every performance was evaluated as to evidence of Whole or Part Approach. Whenever there was evidence of the Part Approach a "D" was recorded. A "W" was recorded when there was evidence of the Whole Approach. If there was only slight evidence of Part or Whole Approach a (D) or (W), respectively, was noted. Then an over-all judgment was made as to Whole-Part Approach. These ranged from "pure Part Approach—D" through "mostly Part Approach—D(W)," "combination Whole-Part Approach—DW," to "mostly Whole Approach—W(D)." A final rating as to Whole Approach was assigned to each individual's performance: a rating of W(D) and a rating of DW were considered "High" in Whole Approach, a rating of D(W), "Medium" in Whole Approach, and a rating of D, "Low" in Whole Approach.

Evidence of Rigidity. (a) Produces only prefixes. (b) Produces mainly words. (c) No production of words or prefixes. (d) No succession of sequence. (e) No manipulation of letters. (f) Number of words produced (4 or less). (g) Does not mix up letters in pile. (h) Does not ask himself questions. (i) Does not set up hypotheses. (j) Does not think of words. (k) Makes same errors. (l) Thinks of earlier large word. (m) Gives up.

Scoring procedure of Flexibility-Rigidity. Each

record was examined as to the presence or absence of behaviors that could be taken as evidence of rigidity as stated above. In order to arrive at a total score of rigidity for each subject, the number of categories showing rigid behavior were added for each individual. Thus, subject No. 1 produced only prefixes, showed no succession of sequence, did not manipulate letters, did not mix up letters in pile, and did not set up any hypotheses. His total score was 5. The total scores were then arranged in a frequency distribution and final ratings were assigned as discussed in an earlier section.

Stencil Design Test II

Description. This test (see reference 1) was selected because it involves overt action in a relatively highly structured perceptual situation. It consists of 20 design cards in black and white, each three inches square, and 24 colored stenciled cards, also three inches square. Twelve of these colored cards are yellow, and 12 are blue. There are 11 stencils and 1 solid card of each color; they are numbered in the order in which they are to be arranged on the table when presented to the subject. All the colored cards are reversible so that no time is lost if a card is laid face down. Both the design cards and stencils are symmetrical to avoid difficulty with right-left or up-down reversals. The object is to superimpose blue cards and yellow cards upon each other to form a design similar to the one on the black and white design card.

For this study only five designs were selected for the experiment proper, namely, numbers 9, 13, 10, 18, 19 (1). These designs were selected to provide the proper difficulty and different possible problem-solving approaches for a superior group of students. Problems 10 and 19 can be solved by the Whole Approach whereby the subject has to analyze the design, i.e., determine which stencils are involved. These two problems cannot be solved easily by trial-and-error or by sheer mechanical approach. In contrast,

problems 9, 13, and 18 could be solved mechanically, by placing stencils of alternate colors on top of each other.

Administration. In the Stencil Design Test the subject is shown a sample card and given the following directions: "Make a design with these stencils so that it will look like the drawing on this card." If the subject does not understand the task, the examiner shows him that the pattern could be copied by placing the blue stencil on the yellow solid card. The stencils are returned to their places. To make certain that the subject understands the directions the examiner asks him to copy design number 3 as a practice exercise (1). If the subject does not yet comprehend the directions he is given design number 5 for further practice. After the subject understands the directions, he is given designs number 9, 13, 10, 18, 19 in that order. The order of presentation is held constant for each subject in order to control any possible influence of one design on the following one. Besides that, this specific order calls for changes in attack so that a mechanical approach to the problems has only a limited chance for successful solutions. The subject is told that there is no time limit on the task. Furthermore, he is encouraged to verbalize what he is thinking, how the problem looks to him and what plans he has to attack the problem. If a subject feels hampered by talking out loud, he is not prodded. Any question by the subject as to procedure, etc., is answered with, "That is up to you."

Movements of stencils and verbalizations by the subject during the performance were recorded by the investigator. A special notation was developed by the investigator to record the movement of stencils. This enabled the investigator to evaluate the responses after the test was completed and did not call for immediate judgment or rating of the performance. As mentioned before, verbalizations were recorded as were pauses, glances at the designs or at the stencils; qualitative observations as, e.g., rapid movements or impulsiveness were noted in the record in proper sequence. In addition, the time of first response, i.e., picking up of a stencil, was recorded as was the total time taken for each problem.

Evidence of Whole-Part Approach. This is one of the more structured situations, and a variety of behaviors was taken as evidence of Whole and Part Responses:

1. *Manipulation of material:* Lack of manipulation of stencils was taken as evidence of a Whole Response while frequent manipulation was taken as a Part Response.

2. *Speed of solution:* Putting stencils together

without much checking back to the design and producing the solution rapidly was taken as a Whole Response, while arriving at a solution after a relatively longer time was taken as evidence of a Part Response (13).

3. *Amount of material included in response:* The following behaviors were considered as evidence of a Whole Response; the individual states relationship of several stencils; he states in inquiry that he looked at the problems as a whole. Evidence of a Part Response: during the inquiry, the individual states that he knew one stencil, color, form; he states that he was looking for a circle, etc.; he notices minor projections, points, etc.; he places stencils one at a time; he divides problems into separate parts; he works out parts separately and then combines them.

4. *Plan of attack:* As a Whole Response, the individual spends a major part of the time on the analysis of the design; he analyzes the problem in terms of stencils needed; he verbalizes a complete plan of attack. As a Part Response, the individual analyzes the problem for each stencil separately.

The relative amount of Whole and Part Responses determines the Whole-Part Approach of the individual.

Scoring procedure of Whole-Part Approach. The performance of each subject was rated on each of the five tasks with respect to Whole-Part Approach in accordance with the criteria set up above. These ratings were added in order to arrive at a total score for each individual. A frequency distribution was made of the total scores, and each subject was given a final rating with respect to Whole Approach on the basis of his position in the distribution.

Evidence of Rigidity. In order to evaluate rigidity of an individual in this test situation several behaviors were considered.

1. The subject makes the same errors and shows perseveration in this way.

2. The subject holds on to the same wrong stencils for more than one response. He does not recognize his mistake and does not discard the wrong stencil.

3. The subject misses cues. He seems to be unable to look at the problem differently.

4. The subject uses the same attack on the problem.

Scoring procedure of Flexibility-Rigidity. Each of the five problems was examined and scored separately as to (a) number of same errors, (b) number of wrong stencils (form) held, (c) number of missed cues, and (d) number of times same attack was made on the problem. These scores were then combined for each problem into a total score by adding up the number of occurrences of these behaviors. The total scores

were then arranged in a frequency distribution and ratings as to rigidity assigned. It should be noted that problems that were not completed by the subject received a rating of "High" rigidity.

In order to arrive at a total estimate of the individual's behavior with respect to rigidity in the Stencil Design Test, the total scores of each of the five problems were arranged in one table. By assigning a score of 5 to "High" ratings, 3 to "Medium" ratings, and 1 to "Low" ratings, a total Rigidity score for the Stencil Design Test was arrived at. Then the total Rigidity scores were arranged in a frequency distribution and final ratings assigned in order to facilitate a comparison between the different tests.

Block Design Test

Description. The Block Design Test (8) is another example of a highly structured test. The four designs used in this investigation were presented in red and white, each three and one-quarter inches square, mounted on a white five-by-eight inch card. Sixteen identical blocks, each a one-inch cube having a side colored with red, white, blue, yellow, blue-yellow, red-white, respectively, were used to reproduce the designs. The designs are drawn in such a way that a block is represented three-sixteenths of an inch smaller than in reality. This size was chosen because it allowed the design to be large enough to be adequately examined, and it made the task more difficult than if the design had been reproduced in its true size.

The changes incorporated by the writer in the Block Design Test cause it to differ from the one used by Kohs (8). One change consisted in leaving out the boundary lines which (in the Kohs set) gave the designs a square outline. It was now more difficult to find a beginning and to analyze the design. All the designs were drawn the same size, in contrast to the Kohs test. This leaves it up to the subject to determine what size he should reproduce. In addition, in order to increase the number of possible approaches, the subject was always given all of the 16 blocks, and it was up to him to decide to use all or only some of the blocks.

The designs used in this investigation varied as to complexity. Problems 1 and 3 represented

designs with clearly defined outlines, while Problems 2 and 4 were more complex. Problem 2 was the only one that had a square outline. In the other designs the figure was not square.

Problems 1 and 3 were also chosen because they lent themselves to a construction with four blocks instead of sixteen blocks.

With respect to Whole-Part Approach it was thought that Problem 3 would be mainly attacked by the Part Approach, while Problem 4 would lend itself more to a Whole Approach. In Problem 3 the figures stand out separately, while Problem 4 is too much of a whole to break it up into any figures.

Administration. In the experiment, the subject was given all the 16 blocks in random arrangement and also the Design 1. He was told, "Make a design with these blocks so that it will look like the drawing on this card." There was no time limit. The subject was again encouraged to think out loud while he was working on the problem.

Movements of blocks and verbalizations by the subject were recorded. In order to record the movements of blocks, the special notation proposed by Rubinstein (12) was used. At times, it became necessary to draw the designs as constructed by the subject when his production became very different from the expected one. Furthermore, pauses and qualitative observations were recorded in addition to verbalizations and moves. Also, the time of first response and total time for each problem were made part of the record.

Evidence of Whole-Part Approach. If the subject approaches the problem as a whole, he may proceed to complete the problem row by row regardless of the design (6). He may verbalize this Whole Approach, or he may show it in his performance, which is usually quick and smooth. If he employs the Part Approach, he may break up the design into parts which he completes separately; or he may complete the problem step by step, checking carefully after each step. Usually, in this approach, the subject follows the pattern or design of the problem.

The relative amount of Whole or Part Responses determines the subject's Whole-Part Approach.

Scoring procedure of Whole-Part Approach. As in the Stencil Test, the performance of each subject was rated on each of the four tasks with respect to Whole-Part Approach in accordance with the criteria set up above. These ratings were added to arrive at a total score for each individual. A final rating for Whole Approach was assigned to each subject on the basis of his position in frequency distribution of total scores.

Evidence of Rigidity. In order to evaluate rigidity of an individual in this test situation

several scores were considered:

1. The subject has difficulty in size perception and does not adjust to the given design.
2. The subject does not change his attack on the problems after failure.
3. The subject abandons part-product although it was correct.
4. The subject gives up trying to solve the problem.
5. The subject jumps from one aspect to another.
6. The subject verbalizes fluidity or lack of organization.
7. The number of errors is high.
8. The subject does not correct errors.
9. The subject repeats errors.
10. The subject holds on to wrong blocks.

Scoring procedure of Flexibility-Rigidity. Each of the four problems was examined and scored separately as to the presence or absence of behaviors that could be taken as evidence of rigidity as stated above. These signs were then added for each subject in each problem.

In order to arrive at a total estimate of a person's behavior with respect to rigidity, the scores of the individual in each problem were added and total score computed. This procedure was used because the range of behavior was not great enough to make a separate rating for each problem meaningful. The total scores were then arranged in a frequency distribution and final ratings assigned as discussed in an earlier section.

Experimental Procedure

The battery of tests was administered to each student individually in two sessions of two hours each. In the first session the Stencil Design Test, Anagram I and II Tests, Block Design Test, and the Function Test were presented in that order. This sequence of problems was believed to avoid boredom and poor motivation by changing tasks from perceptual to verbal ones. This sequence was held constant for all of the subjects.

In the second session, the Rorschach Test was administered and the inquiry conducted. The purpose of presenting the Rorschach Test as the last problem was to make certain that the investigator did not have any bias in the observation of problem-solving behavior because

of information received through the Rorschach Test. Another advantage of presenting the Rorschach Test last was that the subjects did not suspect that any personality variables were being observed. The Rorschach Test could have been detrimental to the experiment because it was known by most of the subjects as a clinical instrument for the evaluation of personal maladjustments and would have aroused anxieties in some of the subjects.

The recording of responses has been discussed in a previous section. It may be worthwhile to point out again that initial response time and total time for the task were recorded, with the examiner using a stop watch.

The examiner sat at the right side of a table observing the movements of the subjects. He recorded moves or verbalizations immediately on sheets of paper. The investigator had developed a shorthand system for the recording of moves in the performance tests. The subject was asked to verbalize thoughts, but was not urged further if it disturbed his performance. The room was a cubicle in the Psychological Laboratory of the Education Department at the University of Chicago. It was possible to work without any interruption.

RESULTS AND DISCUSSION

Whole-Part Approach

Individual Differences

The group showed variability with respect to emphasis on Whole-Part Approach in the Rorschach Test, the Function Test, the Anagram I Test, the Anagram II Test, and the Grace Arthur Stencil Design II Test. The evidence as to individual differences was meager in the Block Design Test. For detailed data, see reference 5.

Consistency of Whole-Part Approach

In order to arrive at a measure of consistency of the Whole-Part Approach, correlations between the various tests were obtained as shown in Table 4. According to Lindquist the value of r re-

TABLE 4
INTERCORRELATIONS OF SIX TESTS WITH RESPECT TO WHOLE-PART APPROACH

Test	Function	Anagram I	Anagram II	Stencil	Block	Total Score Minus Particular Test
Rorschach	.25	.40	.40	.58	.40	.67
Function		.00	-.02	-.02	.10	.08
Anagram I			.25	.53	.36	.48
Anagram II				.29	-.03	.27
Stencil					.62	.66
Block						.48

quired for significance for 19 subjects at the 5% level of confidence is .456, and at the 1% level is .575 (9). The correlations between the Rorschach Test Score and the total scores of the other five tests, and between the Grace Arthur Stencil Design II Test and total scores, yielded a coefficient of .67 and .66 respectively. These correlations are significant at the 1% level. The correlations between the Anagram I Test and total scores, and between the Block Design Test and total scores are both .48. These correlations are significant at the 5% level. The results show that the Function Test measures different behavior from the other tests in the battery. In general, the tests seem to indicate a fair degree of consistency of Whole-Part Approach for the individual, as well as significant differences among individuals.

Flexibility-Rigidity Process

Individual Differences

The group showed variability with respect to Flexibility-Rigidity in the Rorschach Test, Function Test, Anagram I Test, Anagram II Test, Grace Arthur Stencil Design II Test, and Block Design Test (5).

Consistency of Flexibility-Rigidity Process

In order to investigate whether Flexibility-Rigidity is a consistent process, correlations between final Rigidity ratings on the various tests were calculated (see Table 5). The correlations of the single tests with the total score yield results that are significant on the 5% level or better for the Rorschach Test ($r = .54$), the Block Design Test ($r = .54$), and the

TABLE 5
INTERCORRELATIONS OF TESTS FOR 19 SUBJECTS WITH
RESPECT TO FLEXIBILITY-RIGIDITY

Tests	Unstructured Tests			Structured Tests			Total Score Minus Partic- ular Test
	R	A I	F	B	St	A II	
Rorschach							
Anagram I		.51	.42	.26	.17	.10	.54
Function			.19	.00	-.19	-.30	.06
Block Design				.16	.32	.25	.43
Arthur Stencil Design					.83	.34	.54
Anagram II						.50	.58
							.30

Grace Arthur Stencil Design II Test ($r = .58$). These results indicate a fair degree of consistency of the Flexibility-Rigidity Process.

Although these tests seem to correlate to some extent with the total score, it is interesting to note that there is no significant correlation between an unstructured test on the one hand and a structured test on the other. The degree to which a problem is structured is evidently of importance.

The correlation between Flexibility-Rigidity in the Anagram I Test and Anagram II Test is .10. If the content of the test were influential in determining Flexibility-Rigidity scores, these two tests should correlate significantly. There is thus an indication that content enters only to a small degree into the results on Flexibility-Rigidity Process.

This conclusion is reinforced by the examination of the intercorrelations of the Stencil Design Test, the Block Design Test, and the Anagram II Test. The Stencil Design Test and the Block Design Test correlate .83. This correlation is very high and may be in part explained by the similarity of the tests. Both tests include manipulation of materials and analysis of spatial relationships. If the content of the test were the major factor, the tests should not correlate with the Anagram II Test. However, the Stencil Design Test and the Anagram II Test correlate .50, which is significant on the 5% level of confidence. The correlation between the Block Design Test and the Anagram II Test is only .34 and, therefore, statistically not significant. The common factor among the three tests is the structure of the situations. Thus, the correlation between the Stencil Design Test and the Anagram II Test can be explained not on the basis of content but on the basis of structure of the situation.

The Function Test shows an unusual pattern in that it correlates with some of the unstructured and some of the structured tests. This can in part be explained by the differences in structuredness of the four problems comprising the test (e.g., "paper" constitutes an object

with a relatively small degree of structure, while "pliers" represents an object with a relatively high degree of structure).

These correlations imply that individuals who are rigid in the Rorschach Test may also behave rigidly in other unstructured situations. On the other hand, subjects who are rigid in the Stencil Design Test may behave rigidly in other structured problem situations.

Relationship of Whole-Part Approach to Flexibility-Rigidity Process

The investigation of Whole-Part Approach and Flexibility-Rigidity Process raises the question whether these processes are independent of each other. Therefore, product-moment correlations were computed; the results are presented in Table 6.

In order to interpret the intercorrelations it should be remembered that the final ratings for the Whole-Part Approach signified the degree of Whole Approach present, while the final ratings of the Flexibility-Rigidity Process represented the degree of Rigidity a person exhibited in the problem situation.

When the correlations between the Whole-Part Scores and the Flexibility-Rigidity Scores for *each test* were computed, neither the Rorschach Test nor the Anagram I Test showed a relationship between the Whole-Part Approach and Flexibility-Rigidity Process. This indicates that the measures used in these two tests are independent of each other.

The Anagram II Test, the Stencil Design Test, and the Block Design Test showed a statistically significant relationship between scores for the Whole-Part Approach and Flexibility-Rigidity Process (correlations of $-.57$, $-.54$, and $-.66$, respectively). The negative correlations indicate that a person who rates high

TABLE 6
INTERCORRELATIONS BETWEEN FINAL RATINGS OF WHOLE-PART APPROACH
AND FLEXIBILITY-RIGIDITY PROCESS

Whole-Part Approach	Flexibility-Rigidity Process						
	Rorschach	Function	Anagram I	Anagram II	Stencil Design	Block Design	Total Rating
Rorschach	—						
Function	-.03	.03	-.21	-.37	-.24	-.12	-.25
Anagram I	.26	.54	.15	.20	.13	.10	.41
	-.36	-.13	-.32	.00	-.20	-.27	-.34
Anagram II	.09	-.29	-.10	—			
Stencil Design	-.25	-.18	-.15	-.57	.00	.08	-.22
Block Design	.06	.12	.19	-.51	-.54	-.38	-.60
				-.33	-.66	-.66	-.56
Total Rating	-.07	-.13	-.02	-.41	-.41	-.31	-.33

in Rigidity will rate low in Whole Approach. This result is in line with clinical findings on the Rorschach Test.

The results indicate, then, that the unstructured tests show Flexibility-Rigidity Process and Whole-Part Approach as two separate processes, while the more structured tests seem to use similar behaviors as evidence for both Whole-Part Approach and Flexibility-Rigidity Process.

The Function Test shows a correlation of .54 between scores for Whole-Part Approach and Flexibility-Rigidity Process. The Function Test has been shown earlier to be a test that seems to measure different behavior from the other tests.

SUMMARY AND IMPLICATIONS

Review of the Plan of the Study

A battery of six tests involving verbal or perceptual problems, and varying from less structured to more structured, was administered to a group of 19 college students. This group was a highly homogeneous population with respect to age, amount of education, total and subscores in the ACE test, reading comprehension test scores, and writing skill test scores.

The individual test records were analyzed to determine the subjects' scores on

Whole-Part Approach, and on Flexibility-Rigidity Process.

Summary of the Results

Whole-Part Approach

The group showed variability with respect to Whole-Part Approach in all six test situations. Correlations between each test and the total score (see Table 4) were significant for the Rorschach Test and the Stencil Design Test at the 1% level, and for the Anagram I and the Block Design Test at the 5% level. The results indicate a fair degree of consistency of Whole-Part Approach for the individual.

Flexibility-Rigidity Process

The group showed variability with respect to Flexibility-Rigidity Process in all of the six test situations.

The correlations of three single tests with the total score yield results that are significant at the 5% level or better. These results indicate a moderate degree of consistency of the Flexibility-Rigidity Process. However, there is no significant correlation between Flexibility-Rigidity scores on an unstructured and a structured test. On the other hand, correlations among the unstructured tests show

significant correlations, as do also correlations among the structured tests. This would indicate that Flexibility-Rigidity Process is fairly consistent for the individual on tasks that are similar in structure, but not consistent for tasks that differ in structure (see Table 5).

Relationship of Whole-Part Approach to Flexibility-Rigidity Process

The correlations of Whole-Part scores and Flexibility-Rigidity scores among unstructured tests are not significant, while the correlations among structured tests are statistically significant. This would indicate that the unstructured tests show Flexibility-Rigidity Process and Whole-Part Approach as two separate processes, while the more structured tests seem to use similar behaviors as evidence for both Whole-Part Approach and Flexibility-Rigidity Process (see Table 6).

Conclusions

The findings of this study seem to justify the following conclusions:

1. Problem-solving processes can be inferred from observation of problem-solving behavior in special tests.

2. A group of subjects homogeneous as to intelligence shows individual differences with respect to Whole-Part Approach and Flexibility-Rigidity Process.

3. The Whole-Part Approach is a fairly consistent process for the majority of the subjects in situations that differ

as to content and degree of structure.

4. Flexibility-Rigidity is a fairly consistent process for the individual on tasks that are similar in structure, but not for tasks that differ in structure.

5. Whole-Part Approach and Flexibility-Rigidity are two separate processes when observed in less structured situations; in more structured situations the two processes are more closely related.

Implications

Problem solving has mainly been looked upon as purely intellectual behavior. However, the Whole-Part Approach and the Flexibility-Rigidity Process are variables that have often been related to an evaluation of personality characteristics. Thus, an investigation of problem-solving processes may reveal how the individual reacts in certain problem-solving situations, and it might lead to other measures of "functioning or effective intelligence."

A focusing on process makes it possible to make finer differentiations of intellectual functioning among individuals at the superior level of intelligence. The differences are then not due to the fact that a subject solves one more problem than the other subject, but rather to the problem-solving processes used.

The investigation of problem-solving processes also leads to an explanation of why subjects fail to solve certain problems. This information is usually not available if one deals mainly with the end products of the problem-solving activity. One has then to be content to state that the subject cannot solve analogy problems, numerical problems, etc.

It should also be pointed out that the problem-solving processes which have been identified lend themselves to cutting across content categories. Whole-Part Approach and Flexibility-Rigidity are processes that can be identified in problems of a verbal, spatial, or numerical nature.

APPENDIX

BEHAVIORS IDENTIFIED IN EARLIER INVESTIGATIONS AS RELATED TO WHOLE-PART APPROACH

Whole Response

- Sargent (13)—Looking at letters with no attempt to form syllables or combinations.
—A rapid solution in the first few seconds.
- Rorschach (11)—Responds to area as a whole.
- Hanfmann (7)—Formulating hypotheses in thinking without much paying attention to blocks.
—Handling and moving of blocks is reduced to a minimum.
- Rapaport (10)—Passive speculation, with little or no manipulation of blocks.
- Goldstein (6)—To plan ahead ideationally.
—To abstract common properties reflectively, to form hierarchic concepts.
—To hold in mind simultaneously various aspects.

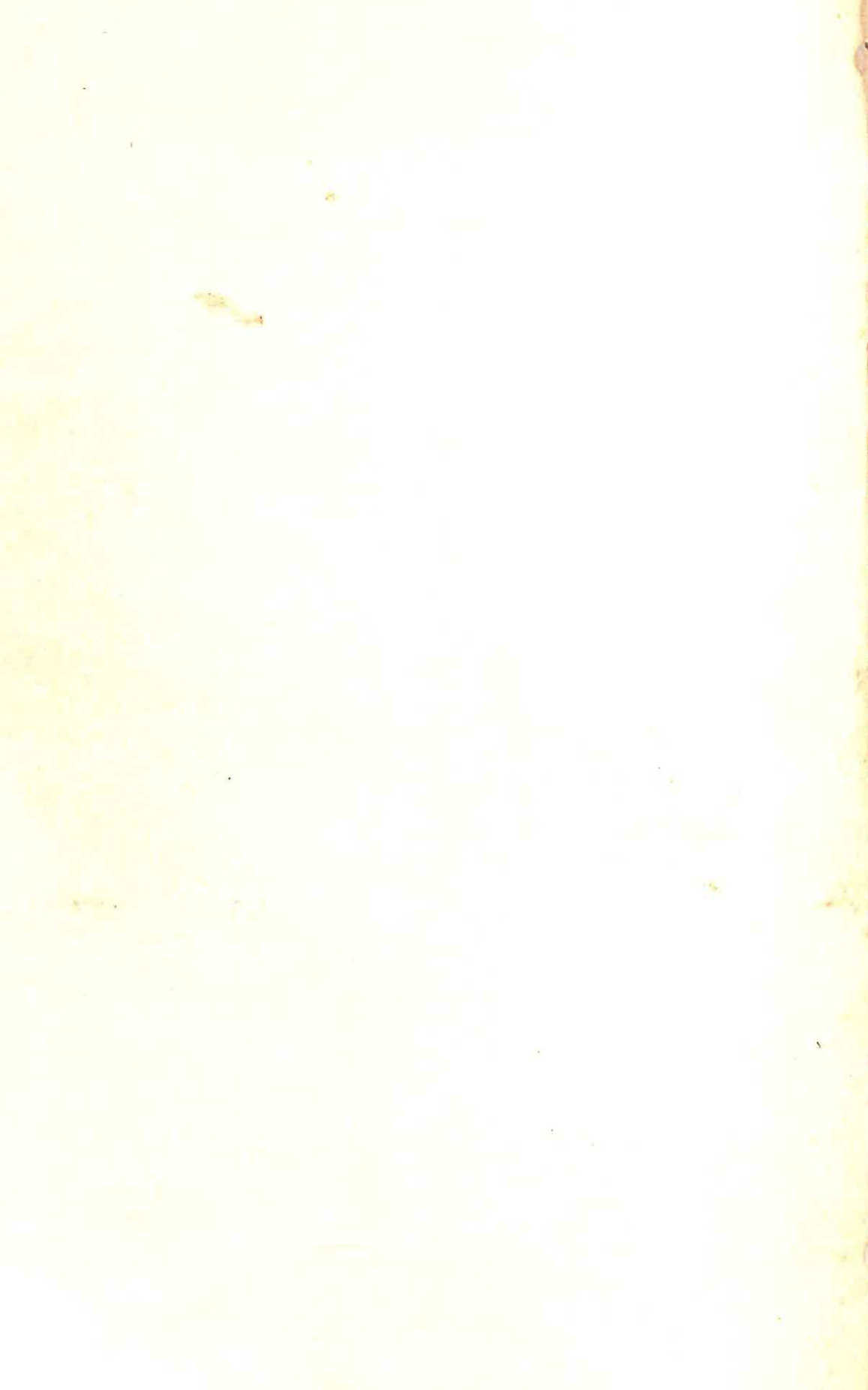
Part Response

- Forming combinations of letters without any definite plan.
—Seeking a common prefix or suffix.
—Making usual consonant-vowel-consonant combinations.
—Trying one letter at a time with all the others.
—Combining letters at random.
—Solution emerges slowly in steps.
- Responds to part of the area.
- Keeps in constant touch with material.
—Gets ideas from looking at blocks and handling them.
—Has groupings first and formulates principles afterwards.
- No attempts are made to organize impressions into generalizations.
- To concentrate on a single aspect of a situation.
- Thinking and acting are directed by the immediate claims which one particular aspect of the object or of the outer-world makes.

REFERENCES

1. ARTHUR, G. *Stencil design test II*. New York: Psychological Corp., 1947.
2. BECK, S. J. *Rorschach's test: I. Basic processes*. New York: Grune and Stratton, 1944.
3. BROWNELL, W. A. "Problem solving," *Psychology of learning*. Forty-First Yearbook of the National Society for the Study of Education, Part II. Chicago: Univer. of Chicago Press, 1942.
4. FISHER, S. Patterns of personality rigidity and some of their determinants. Abstract of unpublished doctoral dissertation, Univer. of Chicago, 1948.
5. GOLDNER, R. H. Individual differences in problem-solving behavior. Unpublished doctoral dissertation, Univer. of Chicago, 1952.
6. GOLDSTEIN, K., & SCHEERER, M. Abstract and concrete behavior: An experimental study with special tests. *Psychol. Monogr.*, 1941, 53, No. 2 (Whole No. 239).
7. HANFMANN, E., & KASANIN, J. Conceptual thinking in schizophrenia. *Nerv. ment. Dis. Monogr.*, 1942, No. 67.
8. KOHS, S. C. *Intelligence measurement*. New York: Macmillan, 1927.
9. LINDQUIST, E. F. *Statistical analysis in educational research*. New York: Houghton Mifflin, 1940.
10. RAPAPORT, D., GILL, M., & SCHAFER, R. *Diagnostic psychological testing: the theory, statistical evaluation, and diagnostic application of a battery of tests*. Vol. II. Chicago: Year Book Publishers, 1946.
11. RORSCHACH, H. *Psychodiagnostics; a diagnostic test based on perception*. Bern: Hans Huber, 1942.
12. RUBINSTEIN, E. A. A note on recording Block Design performance on the Wechsler-Bellevue scales. *J. clin. Psychol.*, 1948, 4, 307-308.
13. SARGENT, S. S. Thinking processes at various levels of difficulty; a quantitative and qualitative study of individual differences. *Arch. Psychol.*, N.Y., 1940, 35, No. 249.
14. WERNER, H. The concept of rigidity: A critical evaluation. *Psychol. Rev.*, 1946, 33, 43-52.

(Accepted for publication March 18, 1957)



Effects of Task Motivation and Expectancy of Accomplishment upon Attempts to Lead¹

JOHN K. HEMPHILL
Educational Testing Service
AND

PAULINE N. PEPINSKY, ARNOLD E. KAUFMAN, MILTON E. LIPETZ
Personnel Research Board, Ohio State University

THE PROBLEM

THIS experiment is the third in a series conducted to test hypotheses about the effects of situational and personal variables upon the frequency of attempts to lead. The first experiment examined the relationship between possession of relevant information and the frequency of attempts to lead (3). The second experiment tested hypotheses concerning the effects of (a) needs for achievement and affiliation, and (b) positive and negative reactions of other group members upon attempts to lead (6). The present experiment tests hypotheses about the effects of two additional variables: task motivation and

expectancy. The hypotheses involved in each of the three experiments are based upon a tentative theory of leadership in small groups (3).

In the tentative theory the following statements relevant to the present experiment appear:

The motivational variables that determine whether a given individual will attempt a given leadership act lie in two major areas. The first area involves disposition variables. Of these variables the most central to our problem include (1) *the strength of his dissatisfactions associated with the mutual problem*, (2) the strength of his social need-dispositions that the group has a potential of satisfying. *The second area involves cognitive estimates or expectations of the individual concerning the probable consequence of the act of leadership he is considering attempting.* Two important areas of variables related to the estimates concern (1) *the probability of solving the mutual problem*, and (2) the probability of increasing or decreasing the potential of the group for the satisfaction of social need-dispositions. We may express the probability of an individual's attempting a leadership act as a function of (1) *his estimate of the probability of the act leading to mutual problem solution*, (2) his estimate of the probable effect of the act on the potential of the group for need-disposition satisfaction, (3) *the degree of dissatisfaction he feels with the mutual problem*, and (4) the strengths of his relevant social need-disposition (3, p. A-26). (Italics do not appear in the original but are used here to emphasize the most relevant statements.)

Three hypotheses, which are related to the above excerpt from the theory, were formulated:

I. *Individuals who perceive that the*

¹The authors wish to thank Delos Wickens for his cooperation in securing subjects. Appreciation is due to the students who participated as subjects in the experiment and to Mrs. Ruth Ann Young, Mrs. Sadie H. Richie, and Mrs. Katherine Mulligan, who rendered invaluable clerical assistance either in the conduct of the experiment or in the preparation of the manuscript.

This research was conducted under contract N6ori-17 T.O. III NR 171 123 between the Ohio State University Research Foundation and the U. S. Office of Naval Research. The opinions expressed are those of the authors alone. This paper is a condensed revision of a technical report bearing the same title submitted to the Office of Naval Research. Sample copies of forms used in the study are appended to that report, which may be obtained on loan from The Ohio State University library.

solution of a mutual problem will be more rewarding attempt leadership with greater frequency than individuals who perceive less reward associated with the solution.

II. *In a situation in which acts of leadership cannot result in a solution of the mutual problem, individuals attempt leadership with less frequency than individuals who are in another situation in which a relationship may exist between acts of leadership and the solution.*

III. *Interaction between these two variables is insignificant in its effect upon the frequency of attempted leadership.*

PROCEDURE

General Design of the Experiment

The experiment involved two independent variables, task motivation and expectancy of accomplishment, and a dependent variable, frequency of attempted leadership acts. Both task motivation and expectancy of accomplishment were varied at two levels. Each of the 24 groups began the experiment under similar expectancy conditions, but one-half of the groups worked under conditions of "high" task motivation, while the other half worked under conditions of "low" task motivation. Midway in the experiment, expectancy conditions were altered so that one-half of the "highly" motivated groups worked under conditions of "high" expectancy, i.e., conditions under which relatively large accomplishment was possible. The second half of the "highly" motivated groups continued the experiment under conditions of "low" expectancy in which, if the situation was analyzed objectively, it was impossible to accomplish the task. The groups working under "low" motivation also were divided into "high" and "low" expectancy groups for the second half

of the experiment. Thus, the experiment was conducted in two sessions, the second following immediately after the first. Session I provided a situation in which the two conditions of task motivation could be established and their effects observed without complication. Session II provided an opportunity to study the effects of both task motivation and expectancy and also their interaction.

The research design also provided for (a) independent checks upon the degree to which the two independent variables actually were manipulated, (b) observation of the frequency of attempts to lead, and (c) control of extraneous factors likely to influence attempts to lead.

The remainder of this section of the report is devoted to detailed descriptions of the selection of the subjects, the Manufacturing Problem, the procedures used to vary task motivation and expectancy, and the procedure used to check upon the establishment of the desired experimental conditions.

Subjects

Subjects were selected from the men enrolled as students in the introductory psychology course at The Ohio State University. Each subject was able to fulfill part of his course requirement by taking part in this experiment. Subjects ranged in age from 18 to 26 years. Care was exercised to avoid including individuals with obviously deviant physical characteristics.

The 96 volunteer subjects were assigned to 24 four-man groups without reference to other criteria than those of scheduling convenience and the assignment of previously unacquainted individuals to the same group.

The Manufacturing Problem

Ten preliminary groups were run before a satisfactory procedure was devised for varying task motivation. The major problem was the creation of a condition of "low" task motivation. This difficulty was due in part to our decision to utilize in this experiment a modification of the Manufacturing Problem (6), which was used as the experimental task in the previous

experiment, and which possesses a high degree of intrinsic interest. Since the procedure for varying task motivation is closely bound to the nature of this task, a general description of it will be given at this point.

The Manufacturing Problem is essentially an elaborate construction task. The task requires each group to organize its members for the purpose of constructing toy models from Tinkertoy parts in a manner to realize maximum credit gains. The groups are called upon to engage in discussion, planning, assembly, reasoning, and arithmetical computation. This task incorporates the characteristics of a number of the tasks frequently used separately in small-group research.

At the outset, each group is given either three dollars in cash ("high" motivation) or 300 points ("low" motivation) with which to begin work. Credit is given only for toys exactly like the models displayed. Each group member is provided with a booklet of information, which itemizes the cost (in terms of either money or points) to the group of separate Tinkertoy parts to be secured from the supplier, and gives the amount of credit to be earned by completing each toy. The subjects are informed that the experiment includes two sessions. It is pointed out that after the first work period both the cost and the credit schedules, as well as the amount to be gained from the construction of the different toys, will be changed. Each 20-minute work session is preceded by a 5-minute planning period; the first work session and the second planning period are separated by a very short break. A clock above the buyer's table is marked off into successive intervals that comprise the two planning periods and the two work sessions.

In order to secure parts, the group is required to fill out special order forms, have them countersigned by all group members, and submit them to the experimenter. After the instructions are read by the experimenter, the organization and operation of the group is left completely in the hands of the members, except that during the designated planning periods no actual construction of toys is permitted. It is emphasized, however, that discussion and revision of plans are not limited to the planning periods but can be engaged in at any time during the work sessions.

The physical arrangements for the task involve the use of four tables placed around the laboratory. These tables are identified by their functions: (a) a supply table, where a large quantity and variety of Tinkertoy parts are available in appropriately labeled boxes; (b) a large table upon which are placed various forms, scratch pads, and pencils, and upon which five completed Tinkertoy models are displayed (a

"top," a "man," an "airplane," a "wagon," and a "ladder"); (c) a table to which finished toys are brought in order to receive credit; and (d) a table upon which copies of current magazines are placed.

In the earlier experiment (6) this task proved to have many desirable characteristics as a *mutual* problem for men college students. Among these characteristics are: (a) a wide variety of interlocked or dependent subtasks permitting organization, division of labor, and utilization of individual differences in ability; (b) high intrinsic interest value, probably introduced by the close similarity to "running a business"; and (c) a direct measure of quality of group performance, as indicated by the amount of profit earned.

Procedure Used to Vary Task Motivation

Under both the "high" and "low" task motivation conditions, the task has the following features:

(a) There are two work sessions, each consisting of a 5-minute planning period and a 20-minute work period.

(b) Tinkertoy parts from which model toys are constructed are secured from the experimenter.

(c) The group members decide which of five toys to construct and how many of each to make.

(d) Completed toys are turned in to the experimenter for credit.

(e) There is opportunity for division of labor among the members of the group.

(f) An account is maintained of the resources ("capital") currently available to the group for the conduct of its operations.

(g) There are no restrictions on free communication among members of the group.

(h) Identical materials are available with which to construct toys, and the arrangement of the laboratory is identical.

The differences between the task as presented under the "high" and the "low" task motivation conditions pertained to (a) the nature of the *reward*, (b) the *attitude* that the experimenter displays toward the task, (c) the *meaning* of the task, and (d) the amount of *routine* involved in securing parts. To facilitate comparison of the conditions, these different procedures are arranged on following page in parallel columns.

The conditions of "high" and "low" task motivation were established by changing only those task characteristics that subjects, in the preliminary runs, reported to be sources of their interest in the task. Care was exercised to provide under both conditions equality of opportunity to attempt leadership insofar as this variable was independent of task motivation.

	High Motivation	Low Motivation
Reward	Subjects worked with money (coins and bills) and retained all the profits their group made.	Subjects worked for "points." No coins or bills were used.
Experimenter Attitude	Experimenter gave the instructions with enthusiasm. He indicated that the experiment was very important and that it involved money supplied by the U.S. Navy. After the experimental instructions were read, he ignored the magazines on the table and proceeded in businesslike manner to his station as "buyer" and "supplier."	Experimenter gave the instructions without enthusiasm. He indicated by side remarks that the experiment probably would be discarded. He suggested that the subjects might find the task dull and uninteresting. After giving the instructions he proceeded to read the magazines.
Meaning	The task was presented as a simulated manufacturing enterprise and as a very good test of individual and group ability.	The task was presented as a Tinkertoy construction task originally designed for use with school children but not appropriate for college students.
Routine	The routine or "paper work" associated with ordering supplies was minimized by supplying parts in lots or "kits." Only one order form was required.	Parts had to be secured separately (no kits). Order forms were required in triplicate.

Procedure Used to Vary Expectancy of Accomplishment

Since the prime objective of the task was to make profit (or gain points), it was possible to vary the expectancy of accomplishment by changing the relationship between the cost of parts and the selling prices of completed toys. Under conditions of "high" expectancy it was possible for group members to make a relatively large cash profit (or many points) but under conditions of "low" expectancy, it was not. However, the success of such a procedure is likely to depend upon whether the "objective" situation is correctly perceived by the subject. Whether accurate perceptions will occur depends in turn upon two factors: (a) the clarity with which the necessary outcome is presented to the subjects, and (b) the strengths of motives that may influence subjects to misinterpret the facts as presented. The results of recent studies (e.g., 1, 2) of the relation between personality factors and perceptual distortion suggest that we might expect to find many failures to perceive the fact that no profit could be made under "low" expectancy and "high" task motivation. If present, this tendency would introduce an interaction effect between the two main variables of the experiment. In order to minimize the possibility of an interdependence of task motivation and expectancy of accomplishment, an effort was made to maximize the clarity

for the subjects of the profit or point-gain situation. Toward this end the experimenter's instructions to the subjects emphasized the importance of determining the possibilities for gains in advance, and illustrated in detail the procedure for determining profits (or point gains) from the cost and price information supplied. Table 1 shows the difference between the points (or money) the subjects surrendered to obtain supplies and the points (or money) the experimenter credited to the group upon the completion of each toy.

TABLE 1
PROFIT GAIN (IN CENTS) OR POINTS GAIN WITH THE CONSTRUCTION OF EACH OF FIVE TOYS UNDER (a) CONDITIONS OF SESSION I; (b) CONDITIONS OF SESSION II, HIGH EXPECTANCY AND LOW EXPECTANCY

Article	Session I	Session II	
		High Ex-pectancy	Low Ex-pectancy
Man			
Top	3	-1	-1
Airplane	7	20	0
Wagon	15	21	-4
Stepladder	32	0	0
	-6	150	0

It can be seen from Table 1 that the group member who correctly interpreted the data supplied him and proceeded accordingly to work on the task could expect to gain some profit or points during Session I. Furthermore, during Session II, if he were working under the high expectancy condition he would note that on three of the five toys a more attractive gain was possible than he had encountered during his first session. Under low expectancy, however, the situation would be accurately perceived as offering no opportunity to gain profit or points by working on the task.

Procedure for Observing Attempted Leadership Acts

A manual, reproduced elsewhere (3), established rules for differentiating attempted leadership acts from other behavior of group members. In brief, observers located behind a one-way-vision mirror made a tally mark in an appropriate column of a prepared form each time a group member made an imperative statement specifying "a change in the immediately present mode of operation of the group members during the process of mutual problem solution" (3, p. A-40).

This procedure of tabulating attempted leadership acts was utilized without modification in the present experiment. Since an estimate of the reliability of agreement between observers had been made on three other occasions (3, 6, 7) and was uniformly high, no further check of this kind was deemed necessary. It is sufficient to state that the median intraclass correlation within pairs of trained observers is above .9 and ranges from .80 to .96.

One further point should be noted. Two observers were used in this experiment. One observer was available only for those experiments run in the morning, the other only in the afternoon. In order to guard against confounding the experimental variables with interobserver differences in the average number of acts they recorded, the experimental design was counter-balanced so that each condition was represented equally in the morning and afternoon runs.

Checks on Task Motivation and Expectancy

Two checks upon task motivation were built into the experiment. The first was provided by the subjects' responses to an item included in a short Post-Experiment Questionnaire. This item was:

How important was it to you that your group do an outstanding job on the task? *a.* Extremely important. *b.* Fairly important. *c.* Not

very important. *d.* Rather unimportant. *e.* Definitely unimportant.

The second check was provided by the ratings of an additional observer (located behind a one-way-vision mirror) who at the end of each five-minute interval rated the groups on six items of task motivation. These items were as follows:

1. How enthusiastic was the group?
2. How hard did the group members work?
3. How tense did the group members appear to be?
4. How important did doing well on the task appear to be for the group members?
5. How intent were group members upon making points or profit to the exclusion of other things?
6. How motivated in general were the members of the group to make points or profits?

Although the observer usually rated *group* rather than *individual* task motivation, he was instructed to rate individuals separately whenever he perceived differences in motivation among the members of the group.

Thus, as a check upon the procedure used to achieve high or low task motivation in the experiment, we have (*a*) the individual's subjective report about his own task motivation and (*b*) the judgments of an outside observer about task motivation.

An item also included in the Post-Experiment Questionnaire was designed to provide a check on the procedures used to establish the expectancy conditions. This item was:

During the second work session, how difficult did you find it to do well on the task? *a.* Impossible to do well. *b.* Relatively difficult to do well. *c.* Don't know. *d.* Relatively easy to do well. *e.* Very easy to do well.

In the experimenter's view his procedure established an unambiguous difference between the high and low probabilities of accomplishment. But there remained a question as to whether the individual subjects under low expectancy would consider themselves to be "doing well" with reference to other criteria than their performance on the official task.

RESULTS

The degree to which the desired experimental conditions were established will be examined first, after which analyses of the effects of the independent variables upon the dependent variable will be presented.

Checks Upon Experimental Conditions

In order to assess the degree to which

TABLE 2

RESPONSES OF 96 GROUP MEMBERS TO THE POST-EXPERIMENT QUESTIONNAIRE ITEM, "HOW IMPORTANT WAS IT TO YOU THAT YOUR GROUP DO AN OUTSTANDING JOB ON THE TASK?"

Response Category	Conditions							
	Motivation (M)		Expectancy (E)		Combined			
	High	Low	High	Low	High M High E	High M Low E	Low M High E	Low M Low E
a. Extremely Important	28	7	24	11	18	10	6	1
b. Fairly Important	20	27	18	29	6	14	12	15
c. Not Very Important	0	12	6	6	0	0	6	6
d. Rather Unimportant	0	1	0	1	0	0	0	1
e. Definitely Unimportant	0	1	0	1	0	0	0	1

experimental procedures produced the desired conditions it is necessary to answer two questions: (a) How successful were we in establishing high versus low task motivation and high versus low expectancy? (b) Did the variations which were introduced for the purpose of creating high and low expectancy interact in any way with the task motivation of the subjects? The answers to these questions are provided by the subjects' responses to the Post-Experiment Questionnaire and the observers' ratings made during the experimental session.

Motivation

As a check upon the manipulation of task motivation, an analysis was made of the responses to the item of the Post-Experiment Questionnaire that concerned the importance of doing an outstanding job on the task. These responses are presented in Table 2.

Chi-square tests were made of the association between experimental conditions and the frequency of the responses in the five categories. Chi square was partitioned (4) into three components: (a) that associated with the high and low motivation conditions, (b) that associated with the high or low expectancy conditions, and (c) that associated with the

interaction of motivation and expectancy. The total chi-square value is 40.97 (significant at the .01 level), which divides into the components 27.64 (significant at the .01 level) associated with motivation conditions, 7.42 (not significant) associated with expectancy conditions, and 5.91 (not significant) attributable to the interaction of the independent variables. The subjects' reports are consistent with the experimentally created motivation conditions.

We note, however, that only 2 of the 48 subjects run under low motivation choose the responses "rather unimportant" or "definitely unimportant." It appears that the 48 subjects cannot be described as having low motivation in any absolute sense, but rather should be considered as being relatively less motivated than the highly motivated group members.

A second independent check upon the creation of high and low motivation conditions is provided by reference to the ratings made by an outside observer.² At

² Two observers were used; one rated groups run in the morning, and the other rated the afternoon groups. The observers were not informed of the specific experimental conditions under which each group they observed was to be run. However, it is likely that they made an early guess concerning these conditions.

the end of each five-minute interval during the experiment an observer rated the group on six items relevant to task motivation. The observer was instructed to rate the entire group as a unit, *but* if any one or more group members could be differentiated from the others, to rate those members as individuals. Although most ratings were made for the groups as units, some individuals were differentiated. When no differentiations are made among group members, the group rating is assigned to each of the four individuals.

A test of the difference between the means (based on the sum of all ratings, i.e., six items rated at the end of each of 10 five-minute periods) of the high and low motivated groups yields a value of 9.42 (significant at the .01 level). Clearly the observers' ratings correspond to experimental variation in motivational conditions.

In order to examine the degree of correspondence between the observers' ratings of a group member's task motivation and the group member's estimates of the importance of the task, correlations were computed between the ratings by the observers and an index based upon the responses by group members to the first item ("importance") of the Post-Experiment Questionnaire. This index was constructed by assigning to the responses to the Post-Experiment Questionnaire values of one to five points (lowest to highest rated importance). Table 3 presents the correlations between each of the six items and the total ratings both by session and for the two sessions combined.

Each of the six items is correlated positively with the subjects' estimates of the importance of doing well on the task. The item that appears to be most strongly related is, "How tense did the group members appear to be?" These correla-

tions are .40 for the first session and .55 for the second. Two items—"How hard did the members work?" and "How intent were group members upon making points or profits to the exclusion of other things?"—tend to be less strongly related to the subjective estimates. During Session II, the correlations between these latter two rating items and the subjects' estimates (.39 and .40) are both significantly less (.05 level) than for the first cited item (.55). Both these items reflect the level of task-oriented activity of the group members. The first item expresses "tension" or anxiety displayed in the situation. Perhaps the difference in the correlation coefficients is to be explained by the behavior of individuals who quit work under high motivation and low expectancy, but who, nevertheless, exhibit other signs of their high motivation. If the item, "How hard did the group members work?" is excluded from the total motivation rating, some of the effect of amount of activity as an indication of motivation may be removed. Then the correlation between the observers' ratings and the subjects' estimate of the importance of the task becomes .65 (instead of .53 for all six items).

These checks on the procedures used to vary task motivation show that (a) the subjects attach less importance to doing an outstanding job on the task under conditions of low motivation than under conditions of high motivation, (b) outside observers rate motivation higher under conditions of high motivation than under conditions of low motivation, and (c) the subjects and the observers tend to agree about the subjects' motivation to work on the task. It can be inferred, therefore, that the desired motivational conditions did exist for the subjects during the experiment.

Expectancy

The experimental conditions of high and low expectancy were introduced during the second work session only. High expectancy was produced by presenting the subjects with more attractive profit or point schedules than those used during the first session. Low expectancy was established by schedules so constructed that the subjects could not realize a

TABLE 3

CORRELATIONS BETWEEN AN INDEX BASED ON RESPONSES OF GROUP MEMBERS TO AN ITEM ON A POST-EXPERIMENT QUESTIONNAIRE REFLECTING TASK MOTIVATION AND RATINGS OF MOTIVATION MADE BY AN OUTSIDE OBSERVER
($N=96$, Mean Index = 3.88, SD Index = .77)

Rating Item Correlated with the Questionnaire Index	Mean	SD	r^*
<i>First Session</i>			
How enthusiastic was the group?	18.2	4.82	.37
How hard did the group members work?	21.5	4.00	.31
How tense did group members appear to be?	14.2	5.84	.40
How important did doing well on their task appear to be for group members?	25.8	5.31	.36
How intent were group members upon making points or profit to the exclusion of other things?	27.7	3.43	.27
How motivated in general were members of the group to make points or profits?	21.0	6.15	.32
Sum of the above items.	128.3	26.42	.40
<i>Second Session</i>			
How enthusiastic was the group?	15.3	9.42	.46
How hard did the group members work?	18.5	8.58	.39
How tense did group members appear to be?	13.7	8.78	.55
How important did doing well on their task appear to be for group members?	18.4	10.18	.47
How intent were group members upon making points or profits to the exclusion of other things?	19.2	9.80	.40
How motivated in general were members of the group to make points or profits?	16.6	9.88	.48
Sum of the above items.	101.6	52.66	.40
<i>Session I & II Combined</i>			
How enthusiastic was the group?	33.5	11.80	.51
How hard did the group members work?	39.9	11.88	.41
How tense did group members appear to be?	27.8	13.35	.54
How important did doing well on their task appear to be for group members?	44.1	13.15	.51
How intent were group members upon making points or profits to the exclusion of other things?	46.9	10.98	.44
How motivated in general were members of the group to make points or profits?	37.6	13.65	.49
Sum of the above items.	229.9	68.04	.53

* A correlation of .27 is significant at the .01 level.

monetary or point gain by working on the task. Were the subjects' perceptions of the conditions consistent with the experimenter's intentions?

Question II on the Post-Experiment Questionnaire reads, "During the *second work session* how difficult did you find it to do well on the task?" *a.* Impossible to do well. *b.* Relatively difficult to do well. *c.* Don't know. *d.* Relatively easy to do well. *e.* Very easy to do well.

This item was deliberately framed to avoid a direct question about "making

profit or points" so that the subject was free to select a *d* or *e* response if he should define "doing well" in terms other than points or profits. The possibility was anticipated that subjects might redefine the task of the group and continue to work for other motives than to make profit (or gain points).

Table 4 represents the distribution of responses to this item of the Post-Experiment Questionnaire.

Chi-square tests were made of the association between the experimental con-

ditions and the frequency of the subjects' responses to each of the five categories. Chi square was partitioned into three components: (a) one associated with the high and low task motivation conditions, (b) a second associated with the high and low expectancy conditions, and (c) the third associated with the interaction of the motivation conditions and the expectancy conditions. The total chi square was found to be 29.45 (significant at the

The Effects of Task Motivation and of Expectancy Conditions Upon the Frequency of Attempted Leadership Acts

Two analyses of variance were performed in order to study the effects of task motivation and expectancy upon attempts to lead. Before these results are presented, descriptions will be given of the transformation procedure employed to convert the observed frequencies of

TABLE 4

DISTRIBUTION OF THE RESPONSES OF 96 GROUP MEMBERS TO THE POST-EXPERIMENT QUESTIONNAIRE ITEM, "DURING THE SECOND SESSION, HOW DIFFICULT DID YOU FIND IT TO DO WELL?"

Response Category	Conditions							
	Motivation (M)		Expectancy (E)		Combined			
	High	Low	High	Low	High M High E	High M Low E	Low M High E	Low M Low E
a. Impossible to do well	9	7	1	15	1	8	0	7
b. Relatively difficult to do well	9	4	8	5	6	3	2	2
c. Don't know	1	7	3	5	0	1	3	4
d. Relatively easy to do well	18	21	26	13	10	8	16	5
e. Very easy to do well	11	9	10	10	7	4	3	6

.01 level), which divides into a component 17.80, associated with the expectancy conditions (significant at the .01 level), a component 7.12, associated with the motivation conditions (not significant), and a component 4.53, associated with the interaction of motivation and expectancy (not significant).

It should be noted that 23 of the 48 low expectancy subjects indicated that it was either "very easy to do well," or "relatively easy to do well." In view of the objective conditions under which they worked, it appears that these subjects must have defined "doing well" in terms other than points or profit gain. In general, however, the responses of the subjects to this item are in accord with the requirements of the experiment.

attempted leadership acts into "attempted leadership scores," and of the pattern of the analyses. After the general plan of analysis has been made clear, the results of the two analyses of variance will be presented. Finally, certain further analyses will be made of the more significant effects revealed by the main analyses.

Attempted Leadership Scores

The data of the attempted leadership scores are the observers' tallies of the number of attempted acts of leadership during specific intervals of time. The frequencies of the occurrence of an event observed within a limited time interval approximate a Poisson distribution. The square root transformation has been recommended (5) for the purpose of stabilizing the variance of these types of data. Since homogeneous within-group variance is required for appropriate use

TABLE 5
DISTRIBUTION OF RAW AND TRANSFORMED^a ATTEMPTED LEADERSHIP SCORES
(N=96)

Session I			Session II		
Raw Score	Transformed Score	Frequency	Raw Score	Transformed Score	Frequency
41-	9	1	41-	9	0
32-40	8	2	32-40	8	2
25-31	7	9	25-31	7	5
18-24	6	8	18-24	6	8
13-17	5	12	13-17	5	6
8-12	4	25	8-12	4	19
5-7	3	19	5-7	3	18
2-4	2	12	2-4	2	23
1	1	5	1	1	8
0	0	3	0	0	7

^a These transformed scores are in each case the greatest integer less than $\sqrt{2x}$; e.g., any value between 7.00 to 7.99 received the value 7.

of analysis of variance, "attempted leadership scores" were computed based on a square root transformation of the frequency data.

Table 5 presents the separate distributions of the original and the transformed ($\sqrt{2x}$) data for Sessions I and II.

Chi-square tests of the departure of the distributions of transformed scores from normal were made for Session I and Session II. The obtained values of chi square are 5.23 for Session I ($p > .5$) and 9.13 for Session II ($p > .2$).

A basic assumption underlying the analysis of variance is that of homogeneity of within-class variance. Bartlett's test was applied to the variances of the attempted leadership scores within classes composed of the four individuals included in each of the 24 groups. Since for each session the probability of the occurrence of the obtained chi-square value is greater than .5, this assumption is met satisfactorily.

Patterns of Analysis

It has been noted that the general design of the experiment made provision for the analysis of the effects upon attempted leadership of two independent variables, task motivation (M) and expectancy of accomplishment (E). The counterbalanced arrangement of the experiment also provides for the evaluation (and removal from experimental error) of the effects of differences between observers (O).

Session II attempted-leadership scores can be analyzed in more detail than Session I scores. Of the available 95 degrees of freedom, 23 are assignable to differences between groups and 72 to differences among individuals within the same group. The 23 degrees of freedom between groups are allocated among the four experimental treatments (i.e., high motivation and high expectancy, high motivation and low expectancy, low motivation and high expectancy, low motivation and low expectancy). One degree of freedom is allotted to each of three effects, motivation, expectancy, and the interaction of motivation and expectancy. Allocations also are made from group differences to provide an evaluation of differences between the two observers who tallied the leadership acts and the interaction of the observer with the expectancy and motivation variables. The remaining 16 degrees of freedom depend upon unique differences among groups, i.e., uncontrolled variation *not* attributable to the effects of motivation, expectancy, observer differences, and their interaction. Expectancy conditions

TABLE 6

ANALYSIS OF THE VARIANCE OF ATTEMPTED LEADERSHIP SCORES FOR SESSION I

Source of Variation	df	Variance	F
Between Motivation Conditions	1	20.17	6.79*
Between Observers	1	10.67	3.59
Interaction; Motivation X Observers	1	2.03	.68
Group Uniqueness	20	4.78	1.61
Within Groups (Error)	72	2.97	
Total	95		

* Significant at the .05 level.

were not varied during Session I and the analysis is correspondingly simplified.

Analysis of Variance

The analysis of the Session I scores provides an evaluation of the task motivation variable uncomplicated by variation in expectancy. The analysis of Session II scores permits an examination of the effects of expectancy, task motivation, and their interaction.

Table 6 presents the results of the analysis of the attempted leadership scores of Session I.

For Session I the effect of the motivation conditions is significant at the .05 level. Expectancy conditions were not varied during Session I and therefore do not appear in this analysis. The F value of 1.61 associated with group uniqueness does not reach the .05 level of significance, and we may assume that the groups do not differ with respect to attempted leadership during the first session, except for the effect of task motivation.

Table 7 presents the analysis of the variance of the attempted leadership scores for Session II.

The analysis of the variance of attempted leadership scores, summarized in Table 7, shows that the differences between observers contributed to the

variation of the scores for Session II. We also note a significant interaction between observers and expectancy. This finding contrasts with the failure of the observer differences to appear as a significant source of variation during Session I (Table 6).

Both task motivation and expectancy are significant sources of variation in attempted leadership scores for Session II. The interaction between expectancy and motivation does not reach significance. Apparently, the two experimental treatments have independent effects upon the attempted leadership scores.

The between-observer variance may be attributed to either (a) differences between the observers' standards or (b) differences in the subjects' behavior associated with the time within the day at which the experiment was conducted. It has been mentioned that one observer was employed during the morning run (8-10 a.m.), and the other worked only in the afternoon (1-3 p.m.).

TABLE 7

ANALYSIS OF THE VARIANCE OF ATTEMPTED LEADERSHIP SCORES FOR SESSION II

Source of Variation	df	Variance	F
Between Expectancy Conditions	1	33.84	11.87**
Between Motivation Conditions	1	14.26	5.06*
Between Observers	1	33.84	11.87**
Interaction; Expectancy X Motivation	1	.10	.04
Interaction; Expectancy X Observers	1	12.76	4.48*
Interaction; Motivation X Observers	1	1.76	.62
Interaction; Expectancy X Motivation X Observers	1	.02	.01
Group Uniqueness	16	3.42	1.20
Within Groups (Error)	72	2.85	
Total	95		

* Significant at the .05 level.

** Significant at the .01 level.



TABLE 8

AVERAGE ATTEMPTED LEADERSHIP SCORES
ASSOCIATED WITH TWO OBSERVERS FOR
TWO CONDITIONS OF EXPECTANCY

Observer and Condition	Average Score
Observer A and High Expectancy	3.63
Observer B and High Expectancy	4.08
Observer A and Low Expectancy	1.71
Observer B and Low Expectancy	3.63

We have one bit of evidence which suggests that diurnal effects (perhaps related to the subjects' alertness in the morning) is the better explanation. Table 8 shows the average attempted leadership scores based on the tabulations of the two observers for the two conditions of expectancy. The obvious discrepancy in Table 8 is the low average attempted leadership score for Observer A under conditions of low expectancy. This observer was employed for the 8-10 a.m. runs.

If the subjects who were run in the morning sessions were more alert than those who were run in the afternoon sessions, they may have perceived the objective conditions of low expectancy more clearly (i.e., realized that it was impossible to make profit or point gains). Examination of the Post-Experiment Questionnaire responses to the item, "During the *second work session*, how difficult did you find it to do well?" reveals that 15 of the 48 group members who experienced the low expectancy condition chose the response, "Impossible to do well," (i.e., their response was entirely consistent with the objective facts). Twelve of the 15 members giving this response were from groups run in the morning sessions. In other words, 50% of the 24 subjects who participated in the experiment in the morning perceived the situation correctly as compared with only 12.5% in the afternoon. The difference between these percentages is significant at the .01 level. It may be, therefore, that the obtained variance between "observers" indirectly reflects the effect of time of day upon the subjects' responses to the expectancy conditions.

In summary, the results show that the desired experimental conditions were established and that the hypothesized relationships are reflected in the data. Task motivation and expectancy of task accomplishment both are significantly related to attempts to lead. It has been demonstrated further that these two vari-

ables operate independently in their effects upon the dependent variable.

DISCUSSION

In general, the results of the experiment support the hypotheses that provided the impetus for the present study and support the theoretical statements to which they are related. Individuals attempt to lead more frequently when the rewards for mutual problem solution are relatively high. They also tend to attempt more leadership acts if an attempted leadership act has a reasonable possibility of contributing to mutual problem solution than if attempts to lead cannot accomplish the task.

Even so, a number of specific factors have operated to minimize the evident effects of task motivation and expectancy on the frequency of attempting to lead. In the two immediately following sections these factors will be discussed in detail. Such an examination may facilitate further efforts to specify more precisely the functional relationships between task motivation and expectancy as those variables determine the frequency of attempts to lead.

Factors Related to Task Motivation

The design of the experiment stipulated that two differentiated levels of task motivation be established for the individual, such that failure to solve the mutual problem would involve less dissatisfaction under one condition than the other. The degree to which those individuals who work under conditions of high motivation feel and show dissatisfaction with the situation that confronts them should be greater than that of group members with low motivation. Checks upon the success of the efforts to establish these two conditions indicated that (a) more subjects regarded doing well on the task as extremely important under high motivation than under low motivation and (b) the observers tended to rate the highly motivated subjects as showing more tension and enthusiasm. There is, however, little to support the view that those who worked under the less motivating conditions had low motivation in an absolute sense. Despite all

the changes made in the task instructions in an effort to induce lower motivation, only approximately 40% of the subjects indicated that it was "not very important." Most subjects found sufficient reasons for regarding the task as "relatively" or "extremely" important.

Perhaps more extreme measures could have been taken to lower task motivation. Yet there are questions to be asked about whether the changes that were made in the present "task" may have altered it in more than its incentive value. Could the effects attributed to task motivation be related to differences in other properties of the task?

The differences between the high and low motivation conditions were introduced in an attempt to reduce the intrinsic interest of the task. To review, the following changes were made:

- a. Identification of the task as a "Tinkertoy construction test" as compared with a "small manufacturing enterprise."
- b. The use of "points" as compared with actual money in the transactions involved in doing the task.
- c. Accumulation of points as the reward for accomplishment as compared with retention of actual monetary profits.
- d. Requiring that orders for parts be submitted in triplicate as compared with requiring only one order form.
- e. Requiring that parts be ordered separately as compared with permitting orders to be placed for "kits" of parts.

The last two changes (*d* and *e*) were introduced to vary the "routine" character of the task by making it repetitive and boring for the subject. At first glance, it would appear that these two changes might restrict the opportunity for attempted leadership. Observations of the group in action, however, did not bear out this surmise. Instead of restricting the organizing requirements of the task through involving the group members in purely routine individual activities, these features of the task provided a sphere of work in which attempts to lead were appropriate. No differences were apparent in the opportunities for attempting to lead under the high and low task motivation conditions.

Factors Related to Expectancy

One of the most interesting results of the experiment was the absence of a significant effect upon attempts to lead of the interaction between task motivation and expectancy. It would not have been surprising if, under the low expectancy condition, the subjects working under high task motivation had displayed a disproportionately greater frequency of attempts to lead than those working under low motivation. One could have reasoned that the highly moti-

vated subjects spent a disproportionately greater effort exploring all possibilities before accepting the objective limits of the situation. The facts are, however, that two of the six groups that were run under high motivation and low expectancy quit work before the end of Session II. The remaining four groups contained at least one member who insisted on working until the end of the period.

Under the low expectancy condition, reasons for continuing work were often discussed by the group members. Such comments as, "We can't lose," or "We're supposed to do something," were made frequently by one or more subjects. It seemed incredible to some subjects that they should encounter a situation in which hard work could not accomplish the task. Much time was spent in the early part of the session in checking and rechecking the available information. In fact, many (more than one-third) of the acts of attempted leadership were made before all the members of the specific groups had realized that nothing was to be gained by trying.

Individuals could—and frequently did—make errors of reasoning that resulted in their drawing erroneous conclusions about the possibility of accomplishment. If it had been feasible to eliminate from the analysis all attempted leadership acts made when their authors believed that a profit was possible (even though it was not), the effect of expectancy on attempted leadership would have been even more striking.

The subjects began the experiment (Session I) under expectancy conditions that proved to be similar to the high expectancy conditions of Session II. The fact that larger rewards were possible during Session II did not appear to have a noticeable effect upon most subjects. Only rarely did a group member comment upon the increased attractiveness of the profit situation. The number of leadership acts attempted during Session II under conditions of high expectancy was not significantly larger than the number attempted during Session I.

There remain many problems of specifying more precisely the relationships between the frequency of attempts to lead and (*a*) the objective characteristics of a situation, (*b*) the perceptions of these characteristics by members of the groups, (*c*) the related expectancies of problem solutions, and (*d*) the degree of certainty associated with these expectancies. The present results are limited in showing only that major differences in the objective situation do affect the frequency of attempted leadership acts.

Pre-Established Structures-in-Interactions

During Session I many groups were successful in establishing patterns of organization that persisted with varying consequences into Session II.

It was possible for the group members to decide upon set ways to perform the task because it remained unchanged in the second session except for the introduction of a new profit or point schedule. The main activities of members continued to be to secure parts from the experimenter by using the order forms, to assemble the same kinds of toys, and to return finished articles to the experimenter for credit. If, for example, the members of a group had worked out a seemingly efficient division of labor for the production of "tops" during Session I, there was a tendency for them to continue with the same organization and to build "tops" during Session II.

Such a consistency in the behavior of a group is referred to in the present theory (3) as a "pre-established structure-in-interaction." It is also suggested in the theory that these structures may replace leadership behavior, but at a cost of flexibility in the solution of group problems (3, pp. 17, 32). Both circumstances were noted in informal observations of the experimental groups. Some groups continued to build the same toy in the second session that they had built in the first, even after they commented upon a disadvantage in possible gain or profit.

Pre-established structures-in-interaction worked against an unambiguous outcome of the experiment in two ways: (a) a decision after the first session to continue to work in the same way during the second session obscured the expectancy variable by diverting the attention of the group members from other more favorable alternatives, and (b) the number of leadership acts that might otherwise have been attempted during the second session was reduced by a single decision simply to adhere to the prior arrangements.

These observations concerning the function of pre-established structure-in-interaction, although consistent with the theory, serve only to suggest additional controls that might have been employed in the experiment. A single substitution of different models of toys to be built during Session II might have reduced the tendency to persist in pre-established organization.

Summary of Discussion

The results of the experiment support the hypotheses drawn from the theory. Task motivation and expectancy are but two of the important variables that must be explored if we are to understand what prompts an individual to try to lead his group. Although we were relatively successful in our efforts to induce variation in task motivation and in expectancy, improvement in the research design might readily produce more conclusive results. However, what

is perhaps the next step in the investigation of these problems is to begin to specify more exactly what properties of tasks are most responsible for the observed effects upon attempted leadership. In this way it may be possible to predict individual attempts to lead, from knowledge of the motivational and expectancy characteristics of various situations.

FINAL SUMMARY AND CONCLUSIONS

Summary of the Experiment

This experiment, the third of a series concerned with attempted leadership, was designed to test hypotheses concerning the effects of task motivation and expectancy of accomplishment.

During the first 25-minute session, the experimental groups were run under either high or low task motivation but constant expectancy conditions. During the second session, four experimental conditions existed, namely, (a) high task motivation and high expectancy, (b) high task motivation and low expectancy, (c) low task motivation and high expectancy, and (d) low task motivation and low expectancy. The data obtained during the first session provided for an uncomplicated test of the effects of task motivation on the frequency of attempted leadership. The effects of both variables and of their interaction could be studied by analysis of the data collected during the second session.

The task used in the experiment was a modified form of the Manufacturing Problem. It involves constructing toy models from Tinkertoy parts obtained from the experimenter in exchange for points or money with the object of making maximum gains from the manufacture and sale of toys to the experimenter.

Twenty-four groups of four men each were run during the experiment. Under each of the four experimental conditions six groups were assigned in a counter-balanced order to morning and after-

noon runs so that interobserver differences could be removed from the statistical tests of the effects of the independent variables.

The dependent variable was the frequency of attempted leadership acts. A tally was made by a trained observer each time a group member attempted to lead his group. As the first step in the analysis, these frequency data were transformed into scores appropriate for the analysis of variance.

The results of the main analysis tended to confirm each of the three hypotheses. In each of the two experimental sessions the effect of task motivation was significant at the .05 level. Expectancy, varied during the second session only, was significant at the .01 level. The interaction of expectancy and task motivation was not significant.

The major findings have been discussed with reference to informal observations that suggest ways of improving and extending investigation of the variables. Task motivation is difficult to establish at a low level, and procedures designed to manipulate this variable may have unanticipated effects upon other properties of the task. Further work is needed in order to clarify the operation of the variables involved in task motivation. The effects of expectancy may have been partially obscured by events inter-

vening between the objectively defined situation and the individual's interpretation of it. One special type of such intervening event is the early establishment of structures-in-interaction that may influence both the requirements for attempting leadership and the kinds of decisions that are made in the face of the facts of the situation.

Conclusions

In general, the conclusions of the experiment are as follows:

1. Individuals will attempt to lead more frequently if the rewards for task solution are high rather than low.
2. If individuals expect acts of leadership to result in problem solution, they will attempt to lead more frequently than if they see no relationship between leading and solving the problem.
3. In this experiment, task motivation and expectancy did not interact significantly in their effects upon attempted leadership.
4. In the design of further experiments employing these variables, it is recommended that careful attention be given to both the potential effects upon other task characteristics of procedures utilized to vary task motivation and the processes that may intervene between the situation as presented and the situation as perceived.

REFERENCES

1. ASCH, S. E. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh: Carnegie Press, 1951. Pp. 177-190.
2. BRUNER, J. S. Personality dynamics and the process of perceiving. In R. R. Blake and G. V. Ramsey (Eds.), *Perception: an approach to personality*. New York: Ronald, 1951. Pp. 121-147.
3. HEMPHILL, J. K., PEPINSKY, PAULINE N., SHEVITZ, R. N., JAYNES, W. E., & CHRISTNER, CHARLOTTE A. *Leadership acts: I. An investigation of the relation between possession of task relevant information and attempts to lead*. Columbus: Ohio State Univer. Res. Found., 1954.
4. MATHER, K. *Statistical analysis in biology*. London: Methuen, 1943.
5. MOSTELLER, F., & BUSH, R. R. Selective quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology*. Vol. 1.

Theory and methods. Cambridge: Addison-Wesley, 1954. Pp. 289-334.

6. PEPINSKY, PAULINE N., HEMPHILL, J. K., & SHEVITZ, R. N. *Leadership acts: II. The relation between needs for achievement and affiliations and attempts to lead under conditions of acceptance and rejection*. Colum-

bus: Ohio State Univer. Res. Found., 1955.

7. SHEVITZ, R. N. *Leadership acts: II'. An investigation of the relation between exclusive possession of information and attempts to lead*. Columbus: Ohio State Univer. Res. Found., 1955.

(Accepted for publication March 19, 1957)

The Pensacola Z Survey: A Study in the Measurement of Authoritarian Tendency

MARSHALL B. JONES¹

U. S. Naval School of Aviation Medicine
Pensacola, Florida

THE THEORY of authoritarianism is independent of political affiliation. Nevertheless, the principal measure of authoritarianism, the California F Scale, is avowedly a measure of "fascist" potential (1). Thus, *in principle*, the F Scale is compounded of at least two sorts of variance: authoritarianism and sociopolitical conservatism. Clearly, for the study of authoritarianism generally, a measure less specific politically is to be desired. The purpose of this monograph is the construction of such an instrument.

The political particularity of the F Scale hinges on the fact that its items are all expressive of a specifically "fascist" point of view. The items are attitudinal in form and sociopolitical in content. These, of course, are not the only materials from which a personality questionnaire may be constructed. In fact, the indirect approach through social attitudes to the measurement of personal qualities is rather unusual. Much more common are items which deal directly with the subject: his habits, temperament, and reactions. Importantly, the usual item does not concern contemporary social groups and policies. If we are to avoid political particularity in the measurement of authoritarianism, we might well rely exclusively upon direct materials as devoid as possible of sociopolitical reference. The construction effort reported in this study has been dominated by this consideration.

The results of this effort are two test

¹ Opinions or conclusions contained in this paper are those of the author. They are not to be construed as necessarily reflecting the view or the endorsement of the Navy Department.

instruments: the *Pensacola Z Scale* and the *Pensacola Z Survey*. The first of these is a single-score test constructed against the F Scale but composed of materials as "asocial" as possible. Since the word "authoritarianism" has become inseparably associated with the F Scale, we will use the word "*heteronomy*"² and its affines when speaking of the Z Scale. The word "*authoritarian*" and its affines will be reserved for the F Scale. The Z Survey is a five-score instrument. One of these scales is the Pensacola Z Scale itself. The remaining four, labeled Dependency, Rigidity, Anxiety, and Hostility, are expansions of the principal cluster-components of the Z Scale.

SETTING

The Pensacola Z Scale is essentially a measure built from ordinary "personal-ity" items against an attitudinal measure of fascist potential. The Z Scale is not, however, the first measure that fits this description. In 1951 Gough (3) reported the construction of his Pr Scale. This instrument was built from items of the Minnesota Multiphasic Personality Inventory (5) against the A-S (Anti-Semitism Scale of the California Group (1).

² The term "heteronomy" has been used by Riesman (16) in very much the same sense as in this study. Although the present usage implies specific tests, Riesman's understanding of the term can be taken as the conceptual intent of the Z measures.

The Pr Scale was not completely developed; no retest reliabilities, for example, were reported. Nevertheless, the Pr Scale could have been our starting point. Instead, we chose to start from scratch and build the Z Scale. The reason for this choice lay in the content of the Pr Scale. No less than eight of the items of the Pr Scale are virtually identical to one or another F-Scale item. Attitudinal material of a nature specific to fascist potential was not deleted. This, of course, is no criticism of Gough. His intent was not to create a scale free from social content. For our purposes, however, the inclusion of these items would have prejudiced our entire undertaking.

All subjects involved in this work were Naval Aviation Cadets. The cadets are procured from two sources: civilian life and the fleet. Civilian-procured cadets comprise approximately 80% of the total. All cadets procured from civilian life must have had at least two years of college. Fleet-procured cadets must have graduated from high school. All subjects were tested within a week after reporting for duty, when the cadets from civilian life were still in civilian clothes and the fleet-procured cadets still in their enlisted uniforms. All cadets, whether civilian or fleet-procured, must pass the selection battery of the Naval Air Training Program, be single, and between 18 and 25 years of age. All sections of the country are represented in any sample of appreciable size.

The properties of the F Scale in the cadet population are, of course, a matter of particular concern. In the first place, the original 30-item scale was slightly modified. One item, "It is best to use some pre-war authorities in Germany to keep order and preserve peace," was dropped as no longer topical. The form used in this study had, therefore, 29 items.

Means and standard deviations of the F Scale are generally reported in item units. Specifically, the mean or standard deviation is calculated for all 30 or all 29 items, as the case may be. The grand mean is then divided by the number of items involved. In these units the F Scale in the cadet population has a mean approximately equal to 3.90 and a standard deviation approximately equal to .75. These parameters apply approximately to all of the samples studied in this work. Neither the mean nor the variance of the F Scale in the cadet population departs

radically from the values reported in *The Authoritarian Personality*. In a typical sample of 311 cadets the F Scale was distributed³ with a negative skew, $g_1 = .31$, and a slight platykurtosis, $g_2 = -.08$.

The homogeneity of the F Scale in a sample of 291 cadets was found to be .81, using the Hoyt analysis of variance technique (8). This coefficient is somewhat lower than the value reported in *The Authoritarian Personality*.

The 24-hour, test-retest reliability of the F Scale in a sample of 80 cadets was found to be +.90 (6). This result corresponds closely to the results obtained in the original study. The 4-week, test-retest coefficient is considerably lower, +.76. The correspondence of this result to the original work cannot be determined; there was no comparable coefficient quoted in the California study. The drop is not, however, as dramatic as it may seem. In the cadet population, almost all personality schedules with retest coefficients in the high .80's or low .90's at 24-hour retest drop to the mid .70's at 4-week retest.

PREPARATION

Before the Pensacola Z Scale could be constructed, a preliminary pool of items had to be created. In writing these items, we had the choice of taking items more or less at random or of attempting some sort of organization. In the latter event, we could have resorted either to a theoretical or to an empirical basis. Both were used. The studies which follow were conducted so that we might have some idea as to what sorts of items could be expected to relate to the F Scale. These relations might then provide an empirical basis for the writing of items. Altogether there were *four studies*. In each of these, a standard personality inventory was administered in conjunction with the F Scale. The four inventories were: the Taylor Manifest Anxiety Scale (17), the Wesley Manifest Rigidity Scale (20), the Guilford-Zimmerman Temperament Survey (4), and the Thurstone Temperament Schedule (18). All four measures had been developed and

³ For a precise definition of the statistics, g_1 and g_2 , see McNemar (15, p. 27).

TABLE 1

CORRELATIONS OF THE F SCALE WITH THE TEN SCALES OF THE GUILFORD-ZIMMERMAN TEMPERAMENT SURVEY AND WITH THE SEVEN SCALES OF THE THURSTONE TEMPERAMENT SCHEDULE

Scale	G ^a	R	A	S	E	O	F	P	T	M
Scales from the GZTS	.10	-.10	-.01	-.08	-.12	-.22	-.24	-.27	-.05	-.07
	A	V	I	P	E	S	P			
Scales from the TTS	.04	.13	.22	-.01	-.17	.23	-.02			

^a Each capitalized letter denotes the scale of the GZTS or of the TTS which begins with the letter involved.

were intended for use in "normal" populations. The principal consideration that prompted this restriction was that the Z Scale was designed for use in normal populations.

In two samples of 166 and 245 cadets, respectively, rank correlations (14) of $+.13$ and $+.14$ were obtained between the F Scale and the Taylor Manifest Anxiety Scale.⁴ The first coefficient is significant beyond the .02 level⁵ and the second beyond the .01 level. In the second study a rank correlation of $+.19$ was found between the F Scale and the Wesley Manifest Rigidity Scale in a sample of 221 cadets. This result is significant beyond the .001 level. In the third study the Guilford-Zimmerman Temperament Survey was administered to a sample of 628 cadets together with the F Scale. The ten scales of the GZTS are: General Activity, Restraint, Ascendancy, Sociability, Objectivity, Friendliness, Personal Relations, Thoughtfulness, and Masculinity. Of these, four had correlations with the F Scale which were significant beyond the .01 level. Two of the four, Emotional Stability and Objectivity, are heavily loaded with anxiety items, an interpretation which is supported by the correlation of $+.66$ which obtains between them. The other two scales, Friendliness and Personal Relations, also enjoy a heavy common factor, the correlation between them being $+.50$. In this case hostility items appear to be the basis of the correlation. In the fourth study the Thurstone Temperament Schedule was paired with the F Scale in a sample of 304 cadets. The seven scales of the TTS are: Active, Vigorous, Impulsive, Dominant, Emotionally Stable, Sociable, and Reflective. Of the seven scales, four had correla-

tions with the F Scale which were significant beyond the .01 level. The first of these, Emotionally Stable, is, like Emotional Stability in the GZTS, essentially an anxiety scale. Its relationship with the F Scale constitutes, therefore, the fourth positive relationship between anxiety and F-authoritarianism. The second scale to relate significantly to the F Scale was Sociable. This scale, like its counterpart in the GZTS, Sociability, is largely a measure of social extraversion. Sociability, however, did not relate significantly to the F Scale. The third significantly related scale of the TTS was Impulsive. The Restraint Scale of the GZTS corresponds, inversely, to the Impulsive Scale of the TTS. In the third study Restraint was related at the .05 level to the F Scale, and in the consonant direction.

All in all, the four studies yield a remarkably stable portrait of the F-Scale authoritarian. The principal features of this portrait are anxiety, hostility, rigidity, social extraversion, and impulsiveness.

CONSTRUCTION

Of the five traits established by the preparatory studies as authoritarian only three, anxiety, hostility, and rigidity, were deliberately utilized in the writing of items. The principal reason for excluding social extraversion and impulsiveness was that their relationships to the F Scale were both weaker and less consistent than those of the three included traits. A second consideration was that in the author's judgment the two excluded traits seemed much less cen-

⁴ For a fuller account of the preparatory studies see the original reports (11, 12, 13).

⁵ All significance levels quoted in this study are two-tailed.

tral than the three included traits to the authoritarian syndrome. This second "consideration" was, of course, entirely subjective. The preparatory studies, then, resulted in three core traits toward which the items of the Z Scale could be oriented: *anxiety*, *hostility*, and *rigidity*. To these three a fourth, *dependency*, was added without empirical justification. Logically, dependency is the essence of the heteronomous syndrome. Unfortunately, no adequate, objective scale of dependency exists. Dependency, therefore, was added on purely a priori grounds. In sum, then, the items of the Z Scale were oriented toward four core traits: anxiety, hostility, rigidity, and dependency. In writing the items, every effort was, of course, made to exclude sociopolitical reference insofar as possible.

Altogether 200 items were created.⁶ All were written in forced-choice form. The principal consideration which led to adoption of the forced-choice form was the relative unfakability of forced-choice items. The F Scale itself is subject to faking (2). The forced-choice form seemed, on available evidence, to offer good promise of protection against faking. To what extent these hopes were realized is a matter which will be taken up later. Of the 200 items, 50 were oriented toward anxiety, 50 toward hostility, 50 toward rigidity, and 50 toward dependency. The "orientation" of the items was, of course, largely subjective since the author's judgment intervened crucially between the results of the pre-

paratory studies and the writing of the items. As will be seen, however, this subjective orientation of items based, in the main at least, on earlier objective results, turns out to be 87 $\frac{6}{10}$ % effective in estimating the direction, if any, in which an item will relate to the F Scale. To what extent the classification of items into four categories can be justified is again a question which will be taken up later.

The 200 items were prepared together with instructions and administered to 306 cadets almost immediately after they had reported for duty at the Naval Air Station. The F Scale was administered one day later.

On the basis of their F scores the 306 cadets were broken down into three groups: those who scored in the lower 27 $\frac{6}{10}$ % on the F Scale, those who scored in the middle 46 $\frac{6}{10}$ %, and those who scored in the upper 27 $\frac{6}{10}$ %. Rank correlation coefficients were then obtained for each item against the trichotomized F Scale.

In evaluating any given item, a *triple criterion* was imposed: (a) the item had to relate to the F Scale in the theoretically anticipated direction, (b) less than 85 $\frac{6}{10}$ % of the cadets had to answer the item in the same way, and (c) the significance of the item against the F Scale had to reach acceptable levels (.20 or better). Of the 200 items, 78 were significant beyond the .20 level. Of these, 10 did not relate to the F Scale in the theoretically anticipated direction; to an additional two items a majority greater than 85 $\frac{6}{10}$ % responded in the same way. The remaining 66 items constituted the Z Scale. In the appendix the 66 items of the Z Scale are reproduced.

In Table 2 the results of the item analysis are set forth. In Column 1 the items of the Z Scale are identified by number. The numbers correspond to those in the appendix. In Column 2 the proportion of subjects who gave response "A" is noted; and in Column 3 the level of significance to which the item attained (.20, .10, .05, or .01). In the appendix the classification and the authoritarian response of each item are indicated.

In crossvalidating the Z Scale the 66 items were prepared as in the Appendix and administered to 311 cadets who had taken the F Scale one day earlier. In scoring the Z Scale, unit weight was assigned to each authoritarian response. Thus, the maximum score possible was 66 and the minimum 0.

⁶The items of the Taylor and Wesley Scales, the Guilford-Zimmerman Temperament Survey, the Thurstone Temperament Schedule, and, in particular, a forced-choice test of leadership (9) which was developed by the Personnel Research Branch of the Adjutant General's Office with West Point cadets were all utilized in the writing of items for the Pensacola Z Scale.

TABLE 2
PROPORTION OF SUBJECTS RESPONDING "A"
AND SIGNIFICANCE LEVELS FOR THE 66 ITEMS
OF THE PENSACOLA Z SCALE

Item	Proportion "A"	Significance
1	.66	.20
2	.73	.05
3	.77	.01
4	.33	.05
5	.42	.05
6	.58	.05
7	.20	.20
8	.28	.01
9	.68	.05
10	.78	.05
11	.33	.05
12	.85	.10
13	.67	.20
14	.74	.20
15	.41	.01
16	.61	.20
17	.52	.20
18	.58	.05
19	.73	.10
20	.39	.01
21	.32	.05
22	.63	.10
23	.40	.05
24	.85	.05
25	.56	.20
26	.34	.10
27	.53	.10
28	.52	.20
29	.60	.10
30	.58	.01
31	.19	.20
32	.30	.05
33	.52	.05
34	.66	.20
35	.70	.05
36	.48	.20
37	.60	.10
38	.56	.01
39	.25	.01
40	.82	.10
41	.42	.01
42	.73	.05
43	.23	.01
44	.26	.05
45	.64	.05
46	.77	.01
47	.48	.20
48	.33	.01
49	.50	.20
50	.55	.10
51	.63	.20
52	.47	.01
53	.56	.10
54	.78	.01
55	.41	.01
56	.53	.20
57	.81	.05
58	.39	.20
59	.75	.20
60	.85	.05

TABLE 2 (continued)

Item	Proportion "A"	Significance
61	.72	.20
62	.20	.20
63	.33	.05
64	.69	.20
65	.71	.10
66	.56	.20

The Z Scale upon crossvalidation correlated $+ .43$ with the F Scale. No attempt was made further to refine the Z Scale. One reason for not refining the Z Scale further was that the crossvalidation coefficient of $+ .43$ seemed quite adequate as it stood. A second reason rested on the reduction in the number of items which would in all probability have resulted from further refinement. The Pensacola Z Scale was designed to serve (a) as a terminal measure of personal autonomy and (b) as a seed scale from which the Pensacola Z Survey would be built. The Z Survey was to be constructed about the item clusters, presuming their existence, in the Z Scale. This approach demanded a fairly substantial number of items in each cluster about which to build the independent scales of the Z Survey. Indeed, simply to determine whether or not there were item clusters in the Z Scale demanded that there be more than three or four items in a cluster. Further refinement of the Z Scale might, through a reduction of items, have jeopardized both the determination and the expansion of item clusters within the Z Scale. For these reasons, no effort was made to improve the crossvalidation coefficient of $+ .43$.

To determine the test-retest reliability coefficients of the Z Scale, two distinct samples were used. The first sample, of 187 subjects, retook the Z Scale after 24 hours; the second sample, of 123 sub-

jects, retook the test after four weeks. In the 24-hour sample the test-retest coefficient was $+ .87$; in the four-week sample the value was $+ .74$.

The notion of "fakability" is compounded of two distinct meanings. The first of these is that a nonfakable test is one such that the average subject cannot *improve* his score by faking. The second meaning is that a nonfakable test is one on which the average subject cannot appreciably *alter* his score by faking. The fakability of the Pensacola Z Scale was examined in both senses.

To determine fakability in the first sense the test was administered to two entirely distinct groups. The first group received the customary instructions. The second group was instructed to "beat the test," to put down what they thought was the "best" answer, regardless of its truth when applied to them. To determine fakability in the second sense a single group was used which took the test first under the usual instructions and then again under instructions to fake. The correlation between the two administrations was adopted as a measure of "alteration."

In examining fakability in the first sense the Z Scale was administered to 196 subjects under a set to fake. The crossvalidation sample of 311 cadets was used as the second group. In the crossvalidation sample the Z Scale had a mean equal to 35.31 and a standard deviation equal to 6.85. In the "fake" sample the Z Scale had a mean equal to 36.18 and a standard deviation equal to 5.69. The variances in the two groups were unequal, significance reaching beyond the .05 level. Applying the Behrens-Fisher test (10) for unequal variances to the difference between means, we obtain a value of $d = 1.55$ and a value of $\theta = 44^\circ$. This result falls considerably

short of the .05 level of significance. From these results we may fairly assume that the Z Scale is not fakable in the first sense. The significant difference in variances, however, suggests that some items are being faked. Since the mean does not change, the cadets must be faking some items against, and other items with, the key. The faking, however, of even a few of the items would be expected to produce considerable alteration in the subject's score.

To determine fakability in the second sense the Z Scale was administered to 220 cadets with the usual instructions. One day later, they took the test under instructions to fake. The correlation between the two administrations was $+ .52$. Under a set to fake, therefore, a subject's score on the Z Scale is subject to considerable alteration. Nevertheless, there is still considerable communality between the test under the two sets.

The second meaning of "fakability" as alteration gives rise to an associated problem: to what extent is the alteration of one's score related to intelligence? The sample of 220 cadets utilized in examining alteration was also given the Psychological Examination (ACE) of the American Council on Education (19). Consider, therefore, the correlation between the ACE and the Z Scale under the two sets. If the more intelligent subjects can "fake better" than the less intelligent, then the correlation between the ACE and the Z Scale under instructions to fake should be higher than the correlation between the ACE and the Z Scale when given under the usual instructions. In other words, the correlation between the ACE and the Z Scale when given normally represents the "legitimate" advantage in autonomy of the intelligent. If, in addition, the intelligent can "fake better," then to this basal advantage there should accrue to the intelligent an increment in Z representative of their more effective efforts at dishonesty. To test, therefore, the hypothesis that the intelligent can "fake better" we need only compare the correlations between the ACE and the Z Scale under the two sets.

The ACE actually consists of two scales: the first is a measure of verbal and the second a measure of numerical ability. Since verbal aptitude might be particularly germane to the

problem, two measures from the ACE were used: the verbal score and the total score, which is the algebraic sum of the two included scores. The correlation between the total score on the ACE and the Z Scale as customarily administered was $-.22$; the correlation between the total score and the Z Scale administered under a set to fake was $-.23$. No entirely satisfactory means exists for testing the significance of a difference between correlation coefficients when the coefficients are based on the same subjects. The procedure developed by Hotelling (7) is, however, the best available. This procedure results in a statistic distributed as the F-ratio with 1 and $(N-3)$ degrees of freedom, where N is the number of subjects involved. In addition to the two correlations involving the ACE, the correlation between the Z Scale under the two sets is necessary in order to calculate Hotelling's statistic. The value of this coefficient was, as already reported, $+.52$. The value of F for the difference between the two coefficients involving the ACE does not approach any acceptable level of significance. The verbal score correlated $-.20$ and $-.26$ with the Z Scale, normal and "fake" respectively. This difference by Hotelling's method yields a result which, like the first, falls far short of significance. From these results we may fairly conclude that the ability to fake on the Z Scale is not related to intelligence. Thus, while the attempt to fake may disrupt a subject's score, he cannot expect to do better, even if he is intelligent.

In Table 3 the distribution parameters of the Z Scale in a sample of 766 cadets are presented. The negative skew is significant at the .01 level; the tendency to platykurtosis does not reach significance at the .05 level. The reader will remember that the F Scale also had a negative skew and a slight tendency toward platykurtosis. The distribution of the Z Scale may, therefore, be a fairly immediate consequence of its construction. In general, the tendency to skew increases with the mean. Samples of high heteronomy are more skewed than less heteronomous samples. All in all, the distribution of the Z Scale does not depart radically from normality. The use of normal statistics would seem justified except with results of borderline significance and high heteronomous potential.

TABLE 3
DISTRIBUTION PARAMETERS OF THE
PENSACOLA Z SCALE

Parameter	Value
N	766
\bar{Z}	35.51
σ_2	6.33
g_1	-.49
g_2	-.25

The homogeneity of the Z Scale in the crossvalidation sample of 311 cadets was $+.72$. This value was obtained using the Hoyt analysis of variance procedure. The Z Scale is at least moderately heterogeneous. This result is not, of course, entirely unexpected. In writing the items of the Z Scale, four kinds of items, distinct at least by intention, were created. To the extent that this intent was realized heterogeneity was to be expected. The existence of heterogeneity does not, by itself, confirm the fourfold classification of Z-Scale items. More exact confirmation can hinge only on a precise analysis of the item intercorrelations.

The original pool of items in the Pensacola Z Scale consisted of 200 items written in forced-choice form. Of the 200 items, 50 were intended to be dependency, 50 rigidity, 50 anxiety, and 50 hostility items. Of the surviving 66 items of the Z Scale 19 had been originally oriented toward dependency, 20 toward rigidity, 15 toward anxiety, and 12 toward hostility. If the intent of the author in writing the items of the Pensacola Z Scale were realized in fact, then the item correlations within each of the four groups of items should be higher than the item correlations between the groups. In other words, the dependency items should cluster together; as should the rigidity, anxiety, and hostility items.

To examine this question the Z Scale was administered to 403 Naval Aviation

TABLE 4
AVERAGE CORRELATIONS BETWEEN AND WITHIN
THE FOUR ORIGINAL GROUPS OF THE
PENSACOLA Z SCALE

	D	R	A	H
D	.135	.065	.038	.000
R	—	.082	-.007	.035
A	—	—	.140	.037
H	—	—	—	.072

Note.—The figures on the principal diagonal represent the within group averages.

Cadets. Less than half of these subjects had been included in earlier samples. Tetrachoric correlations within every possible pair (2,145 altogether) of the 66 items of the Z Scale were then obtained. Since every item-response was scored as either autonomous or heteronomous, the signs of the correlations have a uniform meaning. A positive correlation means that the persons who gave the autonomous responses to the two items tended to be the same. A negative correlation means that the persons who gave the autonomous response to one of the two items tended to give the heteronomous response to the other.

The average item correlations within and between groups are set forth in Table 4. As may be inferred from Table 4, the average between-group correlation is +.030. The average within-group correlation is +.109. Even the correlations of the weakest group, hostility, average well over twice the size of the between-group mean. The strongest, anxiety, averages five times the between-group mean. In brief, then, there is good evidence that the four core traits are, in fact, present in the composition of the Z Scale.

EXPANSION

The underlying logic of the Pensacola Z Scale was to organize into a single in-

strument those sources of variance in the F Scale which could be incorporated by orthodox asocial materials. In a similar fashion the Z Scale itself may be fractionated into component parts. The internal constitution of the Z Scale was found to contain four distinct clusters of items. These four clusters are distinguished from one another by their content. They are also distinct statistically. In the Z Survey the four core clusters of the Z Scale will be expanded into independently scored measures.

In constructing the Z Scale the nature of the different sources of variance in the F Scale was a matter of concern in itself. Of particular significance was the possibility that asocial materials might abstract from the F Scale its heteronomous heart, its basic authoritarian dynamic. To know that a person was a high-F scorer was not enough. We had also to know whether the result was mediated by the social content of the F Scale or not. In a similar way, to know that a person is heteronomous is not enough. The manner in which a person arrives at a high-Z score may well be as significant as the fact that he does.

Before turning to the construction process a description of its end result may serve to orient the reader. The Z Survey has five scales. The first and central scale is Heteronomy. The *Heteronomy Scale of the Z Survey* is nothing else than the *original Z Scale*. The first 66 items of the Survey are the items of the Z Scale and they appear in the same order. The remaining four scales are: *Dependency*, *Rigidity*, *Anxiety*, and *Hostility*. These, of course, are the scales derived by expansion of the Z-Scale elements. In a logical sense they are peripheral to the central scale, Heteronomy. Each of the four peripheral scales has

40 items. The total Survey has 197 and not 226 items, as might at first be supposed. There is item overlap between the peripheral scales and Heteronomy. To be precise there are 29 items of Heteronomy which occur in one or another of the peripheral scales. Needless to say, there is no item overlap between the peripheral scales themselves. In describing the process whereby the Z Survey was created the construction of the peripheral scales will be taken up first. We will then consider the question as to what effect, if any, the incorporation of the Z Scale into the Z Survey as Heteronomy has had upon the properties of the original scale.

The Z Survey was constructed from a pool of 300 items. Of the 300 items there were 75 in each classification (dependency, rigidity, anxiety, and hostility). The first 66 items were those of the original Z Scale. Of the remaining 234 items approximately one-third were taken over intact or with very little modification from the first or original item-pool of the Z Scale; the remainder were newly written. Naturally, in writing these items, those items which clustered best in the Z Scale were used as models.

The construction of the four peripheral scales after the items had been created was carried out in six steps. The first of these steps was to administer the second item-pool to 310 cadets. The instructions were the same as for the Pensacola Z Scale. The 300 items were then broken down into the four groups of 75 items each. For each of these groups a complete matrix of tetrachoric correlations was obtained. In every case the items were scored as either autonomous or heteronomous. Thus, in the anxiety set, for example, a positive correlation between two items means that the cadets who gave the anxious response to one of the two items tended to give the anxious response to the other as well. A negative correlation means that the cadets who gave the anxious response to one of the two items tended to give the nonanxious response to the other. Once the intercorrelations within each group of 75 items had been obtained, we were in a position to refine and pick out the best items. From this point on in the process each group of 75 items was treated as entirely distinct. No effort was made to obtain the correlations between groups. The final Dependency Scale, for example, was constructed exclusively from those items

TABLE 5
CONSTRUCTION DATA FOR THE
PENSACOLA Z SURVEY

Item	D	R	A	H
Total 75 (av. interitem r)	.108	.130	.182	.099
Items eliminated	9	5	11	7
Working pool (av. interitem r)	.097	.126	.175	.091
Final 40 (av. interitem r)	.107	.244	.206	.179
Est. homogeneity	.01	.03	.04	.00
Actual homogeneity	.86	.80	.90	.82

which had been written for dependency. Thus, there was no possibility that an item written originally for rigidity, say, would end up in the Dependency Scale. This restriction clearly prevented the full exploitation of the information contained in the complete 300-item matrix. The reasons for not obtaining the complete matrix were almost purely practical. A 300-item matrix involves the calculation of almost five times as many coefficients as does the calculation of four 75-item matrices. The expected yield, on the other hand, did not appear to justify the increase in labor. The second step in the construction of the Z Survey was to calculate the average interitem correlation in each set of 2,775 coefficients. In Table 5 these averages appear.

The third step in the process was to eliminate all items from further consideration to which a majority greater than 85% responded in the same way. In Table 5 the numbers of items in each classification eliminated because of extreme splits are noted. Among the 32 items which were eliminated because of bad splits were 6 which had been included in the Pensacola Z Scale. Inasmuch as the items of the Z Scale had themselves been screened for bad splits, this result may seem anomalous. All, however, of the 6 items had had borderline splits in the analysis for the Z Scale. By chance fluctuations alone a few items with near-elimination splits in the earlier analysis would be beyond the 85% cut-off point in the present analysis. On the other hand, however, none of the 6 items had splits beyond 90% in the majority response. Exceptions for these items could have been made. To be perfectly safe, however, the cutoff was applied without exception.

After the elimination of items with extreme splits the total of 75 items was cut to a working pool in each classification which varied between 64 and 70 items. These reduced pools are the "working pools" of Table 5, where the average interitem correlation for each is tabulated. In all four classifications the average cor-

relation is somewhat reduced. The items with extreme splits tended to be strongly correlated with the other items of their kind. Thus, in eliminating these items a certain amount of homogeneity was sacrificed. Actually, however, items as extreme as those eliminated cannot, despite their correlations, contribute much to the variance of the final scale. The fourth step in constructing the Z Survey was to determine the average item correlation for each of the remaining items with the other items of its working pool. The effect of this procedure is to inform us as to which of the items of a given pool have the most in common with the pool as a whole and which the least.

The fifth step in the construction of the Z Survey was actually to select the final items of the four scales. Three considerations guided this selection: (a) all four scales were to have the same number of items, (b) this number was to be as small as possible consistent with the requirement that (c) all four scales have *before crossvalidation* projected levels of homogeneity equal to or greater than $+ .90$; since the levels projected are based on results in which chance was exploited the levels upon crossvalidation might very well sink below $+ .90$. From pilot explorations the number 40 appeared to be the smallest round number that would meet these requirements. Accordingly, the best 40 items in each classification, i.e., the 40 items with the highest average correlations, were determined. The average correlation of each of these items with the other 39 "major" items was then obtained for all four pools. The average correlation of each of the remaining "minor" items with all 40 major items was also obtained for each pool. Not every one of the major items showed a stronger average correlation than did every one of the minor items. The major 40 items had the 40 highest averages when these averages extended over all the items of the working pool, major and minor alike. Against just the major items a few of the strongest minor items enjoyed higher averages than did a few of the weakest major items. To complete the construction of the Z Survey we had only to adjust for this overlap. In no group were there more than three minor items added and, hence, no more than three major items dropped. In the ap-

⁷ The levels of homogeneity were projected through the Spearman-Brown formula,

$$r_{kk} = \frac{Kr_{11}}{1 + (K - 1)r_{11}}$$

in which the average interitem correlation was taken as r_{11} .

pendix the items of the Z Survey appear, and the classifications are noted. In the third row of Table 5 the average within-group correlations of the four scales are noted. In the fourth row the levels of homogeneity anticipated by application of the Spearman-Brown formula are also noted. These values have a minimum of $+ .90$ for Hostility.

The sixth and final step was to determine how well the levels of homogeneity anticipated from the construction of the Z Survey held up in an independent sample of cadets. Accordingly, the Survey was administered as it appears in the appendix to a fresh sample of 230 cadets. The Hoyt reliability coefficients were then calculated for all four scales. The results appear in the last row of Table 5. As was to have been expected, all four coefficients are lower than the estimated values in the original sample. On the other hand, they are all substantially homogeneous.

Before turning to the systematic properties of the Survey scales, a word is in order about their names. The titles, Dependency, Rigidity, etc., are not intended to describe or predict their relations to external measures. The names are intended only to describe the content of the scale.

To examine the question of retest reliability the Z Survey was administered to a new sample of 218 cadets. All subjects were readministered the Survey after 24 hours. Interestingly enough, the four peripheral scales order themselves in the same way on retest reliability as on internal consistency. The least reliable scale was Hostility, $+ .83$, and the most reliable was Anxiety, $+ .89$. Dependency and Rigidity fell in between these two extremes with coefficients of $+ .85$ and $+ .87$ respectively. These were the only retest data collected on the four peripheral scales; no 4-week data were collected.

The Pensacola Z Scale was not fakable either in the sense of mean change or in the sense that more intelligent cadets can "fake better" than less intelligent cadets. Nevertheless, all four of its com-

ponent clusters are fakable, and in both senses. The manner, however, in which the peripheral scales are faked makes quite clear the reasons for the nonfakability of the Z Scale itself.

In examining the fakability of the Z Scale, two different designs were used. In the first design, fakability in the sense of mean change was examined with two distinct samples of cadets. In the second design, a single sample was used to examine fakability in the sense that the ability to fake was intelligence-related. To examine the fakability of the four peripheral scales only one design, involving two distinct samples, was used. The first sample was the same sample of 230 cadets which was used in examining the heterogeneity and intercorrelations of the Survey scales. This sample, was, of course, administered the Survey under normal instructions. The second sample was a group of 221 cadets who were administered the Survey under a set to fake. The actual instructions were the same as had been used earlier in testing the fakability of the Z Scale.

The means and variances of the four peripheral scales under normal and fake sets appear in Table 6. As far as the mean differences are concerned, statistical evaluation is unnecessary. All four differences are highly significant. The two "conformist" scales, Dependency and Rigidity, are faked toward heteronomy, i.e., they are faked toward dependency and toward rigidity. The two "nonconformist" scales, however, are faked toward autonomy. Anxiety is faked toward non-anxiety and hostility is faked toward non-hostility. This split explains why the Z Scale showed no substantial change in its mean under a set to fake. *One-half of the scale was pulling one way while the second half was pulling the*

TABLE 7

CORRELATIONS BETWEEN THE ACE AND THE FOUR PERIPHERAL SCALES OF THE Z SURVEY UNDER NORMAL AND FAKE SETS

Scale	Normal Set (N = 230)	Fake Set (N = 221)
D	-.22 ^a	-.01
R	-.16	-.03
A	.07	-.15
H	.09	-.07

^a The normal and fake coefficients for Dependency and Anxiety are significantly different at the .05 level.

other. The fact that the Z Scale changed, though without significance, in the direction of heteronomy is also intelligible. There are more dependency and rigidity items in the Z Scale than there are anxiety and hostility items.

The subjects of both samples were administered the ACE. In consequence, the correlations of the four peripheral scales under a normal set and under a set to fake were available. In Table 7 these correlations are set forth. In Dependency and Anxiety the difference between the normal and the fake coefficients is significant beyond the .05 level. At first sight there seems to be no consistency to the coefficients of Table 7. There is, however. The more intelligent cadets under a set to fake consistently move further toward the fake direction than do the less intelligent cadets. In Dependency and Rigidity the more intelligent cadets under a normal set are less dependent and less rigid. Under a set to fake there is virtually no relationship between either scale and the ACE. In consequence, the more intelligent cadets have taken longer strides toward dependency and rigidity than have their less intelligent classmates. They have moved further under a set to fake toward conformity. The more intelligent cadets were slightly more anxious and slightly more hostile than their fellows under a normal set. The "conformist" direction is toward nonanxiety and nonhostility. Under a set to fake, the more intelligent cadets move further toward nonanxiety and nonhostility than do the less intelligent cadets. These results explain why the Z Scale showed no change in its relationship to the ACE under a set to fake. The more intelligent cadets move consistently in the "conformist" direction. The Z Scale, however, is scored so that the changes in Dependency and Rigidity are nullified by the opposing changes in Anxiety and Hostility. By the former pair the more intelligent cadets should become relatively more heteronomous; by the second pair they should become more autonomous. *The two tendencies balance out, and the*

TABLE 6
MEANS AND STANDARD DEVIATIONS OF THE FOUR PERIPHERAL SCALES OF THE Z SURVEY UNDER NORMAL AND FAKE SETS

Scale	Normal Set (N = 230)	Fake Set (N = 221)
D	24.92 6.48	29.72 5.41
R	23.17 6.97	29.12 6.88
A	14.90 7.57	9.43 5.83
H	17.69 6.23	13.92 6.24

TABLE 8
DISTRIBUTION PARAMETERS OF THE FOUR
PERIPHERAL SCALES OF THE Z SURVEY

Measure	D	R	A	H
\bar{x}	24.92	23.17	14.90	17.69
σ_x	6.49	6.97	7.63	6.25
g_1	-.30	-.18	.41	.50
g_2	-.57	-.63	-.58	.23

more intelligent cadets end up approximately as they were originally.

The correlations of the four peripheral scales with the ACE under a normal set are not without significance in themselves. Dependency and Rigidity relate negatively to the ACE while Anxiety and Hostility relate slightly positively. There appears to be a tendency for less intelligent cadets to be more conformist and for more intelligent cadets to be less so.

In the homogeneity sample of 230 cadets the distribution parameters of the four peripheral scales were calculated. The results are set forth in Table 8. As would be expected, the means of the two conformist scales are greater than 20, while the means of the two nonconformist scales are less than 20. The variances range themselves in the same order as the homogeneity levels of the four scales. Again, this was to have been expected. Test variance and homogeneity have a common item parameter: average interitem correlation. The manner in which the four scales are skewed is also typical. The two conformist scales are negatively skewed while the two nonconformist scales are positively skewed. With the exception of Hostility all scales are platykurtic. Hostility tends without significance toward leptokurtosis. There does not seem to be any very obvious explanation of this tendency.

In its development and study the 66 items of the Pensacola Z Scale were ad-

ministered alone. After its incorporation into the Z Survey as Heteronomy the Z Scale became the first 66 items of a 197-item questionnaire. The change in context which this incorporation involved could conceivably alter some of the properties of the Z Scale. The matter does not seem very serious since the order and priority of the Z-Scale items have not been changed. Indeed, all that has changed is a set: the cadet in taking the Z Scale knows he is taking a 66-item inventory; in taking Heteronomy he is taking a 197-item inventory. To make sure, however, some of the basic properties of the Z Scale were examined to see if they generalized to Heteronomy.

Applying the Hoyt analysis to the first 66 items of the Survey we obtain an internal consistency coefficient of $+.69$. This result is only slightly less than the value for the Z Scale itself. We may, therefore, conclude that the homogeneity or, if you like, the heterogeneity of the Z Scale, has not been disturbed by its incorporation into the Z Survey.

In the 24-hour retest sample of 218 cadets the reliability of Heteronomy was $+.85$. This value is only a trifle lower than in the original scale.

The distribution of Heteronomy is also essentially like that of the original Z Scale. In the sample of 230 cadets the mean was 35.53 and the standard deviation 5.78. In the same sample there was a definite negative skew, $g_1 = -.38$, and a slight platykurtosis, $g_2 = -.25$. These results are all typical of the original scale.

In the "fake" sample of 221 cadets Heteronomy had a mean of 35.56. This value is virtually identical to the value which obtained in the homogeneity sample. Heteronomy does not, therefore, appear to change its mean under a set to fake any more than does the Z Scale.

The correlation between the ACE and Heteronomy in the sample of 230 cadets was $-.09$. The correlation between the same two measures in the fake sample of 221 cadets was $-.21$. Even though $-.09$ is an unusually low value for the relationship between heteronomy and scholastic aptitude there is no significance to the difference between it and the fake coefficient of $-.21$. Heteronomy, like the Z Scale, does not appear to be fakable either in the sense of mean change or in the sense that the more intelligent cadet can fake better than the less intelligent.

In Table 9 the intercorrelations of the five scales of the Z Survey are set forth. As is clear from the table, the two conformist scales account for a much larger part of the variance of Heteronomy than

TABLE 9
INTERCORRELATIONS OF THE FIVE SCALES
OF THE Z SURVEY

	Z ^a	D	R	A	H
Z	—	.42	.44	.22	.07
D	—	—	.32	-.11	-.33
R	—	—	—	-.28	-.64
A	—	—	—	—	-.02
H	—	—	—	—	—

^a "Z" denotes Heteronomy.

do the two nonconformist scales. Also worth noting are the negative correlations between the conformist and the nonconformist scales. At first, these relations may seem odd. However, in their interpretation we must remember that a high score on the Anxiety Scale, for example, does not mean that the subject is anxious; it means he describes himself as anxious. Though they may be related, to be anxious and to describe oneself as anxious are two different things. With this in mind, the negative relations may seem less anomalous. A person, for example, who describes himself as rigid would not seem likely to describe himself as anxious too. More broadly, cadets, at any rate, may ascribe to themselves conformist or nonconformist traits but they tend not to ascribe both.

SUMMARY

In the attempt to free the measurement of authoritarian tendency from political particularity, it was resolved to use materials as purely personal as possible. From such materials the 66-item *Pensacola Z Scale* was then constructed. The Z Scale correlates +.43 with the California F Scale, has a 24-hour retest coefficient of +.87, is unfakable and heterogeneous (+.72). The heterogeneity of the Z Scale stems from the presence within it of four distinct clusters of items, which may be described as "dependency," "rigidity," "anxiety," and "hostility" items. In the *Pensacola Z Survey* these four clusters are expanded into four corresponding scales. These four scales, together with the central or basic *Pensacola Z Scale*, constitute the five scales of the Z Survey. The homogeneity levels of the four peripheral scales range between .82 and .90. The 24-hour retest coefficients range between .83 and .89. All four scales are fakable. The directions in which the four scales are faked divide them into a conformist pair, Dependency and Rigidity, and a nonconformist pair, Anxiety and Hostility. The two conformist scales correlate negatively with the two nonconformist scales. Finally, the systematic properties of the Z Scale are not substantially altered by its incorporation into the Z Survey.

REFERENCES

1. ADORNO, T. W., FRENKEL-BRUNSWIK, ELSE, LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality*. New York: Harper, 1950.
2. COHN, T. S. *Factors related to scores on the F Scale*. Unpublished doctoral dissertation, Univer. of Michigan, 1953.
3. GOUGH, H. G. Studies of social intolerance. *J. soc. Psychol.*, 1951, **33**, 237-269.
4. GUILFORD, J. P., & ZIMMERMAN, W. S. *Manual to the Guilford-Zimmerman Temperament Survey*. Los Angeles: Sheridan Supply Co., 1949.
5. HATHAWAY, S. R., & MCKINLEY, S. C. *Manual to the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corp., 1943.
6. HOLLANDER, E. P. The relationship of leader-

- ship choice to authoritarianism in a military setting. *U. S. Naval Ach. Aviat. Med. Res. Rep.*, 1953, Proj. No. NM 001 058.16.01.
7. HOTELLING, H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. math. Statist.*, 1940, **11**, 271-283.
 8. HOYT, C. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, **6**, 153-160.
 9. IZARD, C. E., & ROSENBERG, N. Prediction of peer leadership by forced-choice test under varied experimental conditions. *Amer. Psychologist*, 1954, **9**, 397. (Abstract)
 10. JOHNSON, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
 11. JONES, M. B. Aspects of the autonomous personality: I. Manifest anxiety. *U. S. Naval Sch. Aviat. Med. Res. Rep.*, 1953, Proj. No. NM 001 058.25.03.
 12. JONES, M. B. Aspects of the autonomous personality: II. Intolerance of fluctuation, Part I; III. Manifest rigidity; and IV. Traits from the Guilford-Zimmerman Temperament Survey. *U. S. Naval Sch. Aviat. Med. Res. Rep.*, 1954, Proj. No. NM 001 058.25.16.
 13. JONES, M. B. Aspects of the autonomous personality: V. Traits from the Thurstone Temperament Schedule; and VI. The Pensacola Z Scale. *U. S. Naval Sch. Aviat. Med. Res. Rep.*, 1955, Proj. No. NM 001 108 109.04.
 14. KENDALL, M. G. *Rank correlation methods*. London: Griffin, 1948.
 15. MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.
 16. RIESMAN, D., GLAZER, N., & DENNEY, R. *The lonely crowd*. New Haven: Yale University Press, 1950.
 17. TAYLOR, J. A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285-290.
 18. THURSTONE, L. L. *Manual to the Thurstone Temperament Schedule*. Chicago: Science Research Associates, 1949.
 19. THURSTONE, L. L., & THURSTONE, T. G. *Manual to the American Council on Education Psychological Examination for College Freshmen*. New York: Educational Testing Service, 1947.
 20. WESLEY, E. S. J. *Perseverative behavior in a concept formation task as a function of manifest anxiety and of punishment*. Unpublished doctoral dissertation, State Univer. of Iowa, 1953.

(Accepted for publication March 25, 1957)

APPENDIX

In 29 of the first 66 items and in all of the remaining 131 items, one of the two statements in each item is followed by an upper-case letter in parentheses. The statement so indicated is the heteronomous member of the item, and the item belongs to the Dependency, Rigidity, Anxiety, or Hostility Scale according as the letter is D, R, A, or H. There are 40 items in each scale. The first 66 items constitute the Heter-

onomy Scale (Pensacola Z Scale). Those items among the first 66 which are not part of a peripheral scale are marked with a lower-case letter (d, r, a, or h) which indicates the classification of the item: dependency, rigidity, anxiety, or hostility. The 29 items in the Heteronomy Scale which do belong to a peripheral scale have, of course, the corresponding classification within the Heteronomy Scale.

THE PENSACOLA Z SURVEY

In this test you will find pairs of statements having to do with personal characteristics. One member of the pair is labeled A and the other B. You are to select from each pair the statement that BEST describes you. Then indicate the statement you have chosen by making a heavy black mark between the lines under the letters A or B (but not C, D, or E) on your answer sheet. Consider the example shown below.

1. A) You are attractive.
B) You are strong.

If you think *You are strong* describes you better than *You are attractive*, you would put a mark under B on your answer sheet. Your answer sheet would then look like this:

1. A B C D E
 || || || || ||

If you marked B on your answer sheet, it would not necessarily mean that you are extremely strong or that you are not attractive. It would mean that ON THE WHOLE, *You are strong* describes you better than *You are attractive*.

Be sure that you select *one* statement from every pair. You are not permitted to omit any pair of statements. Mark your answers on your answer sheet starting with number 1 and continuing through number 66. You should finish the test in approximately 15 minutes.

1. A) You are too friendly for your own good. (r)
B) Your opinions are often incorrect.
2. A) Taking advantage of a person sexually makes you feel bad.
B) You have no scruples in sex. (H)
3. A) You are anxious. (a)
B) You are conceited.
4. A) To you life is a jungle. (A)
B) To you life is a bowl of cherries.
5. A) You day-dream politically.
B) You don't formulate opinions about issues over which you have no control. (d)
6. A) In political activities you confine your efforts to group action. (d)
B) In political activities you frequently indulge in individual endeavor.
7. A) You like a tightly organized group. (d)
B) You like a loosely organized group.

8. A) You haven't made any mistakes in your life.
B) You can't get the mistakes you have made out of your mind. (a)
9. A) There are some people you could never feel for. (H)
B) Sometimes you feel a real compassion for everyone.
10. A) You like instructions to be specific. (d)
B) You like instructions to be general.
11. A) You are sexually appealing.
B) You are faithful. (D)
12. A) You are responsible for most of your troubles.
B) You sometimes get confused without any reason. (a)
13. A) You frequently laugh at yourself.
B) You don't like your favorite habits ridiculed. (r)
14. A) You frequently get away with murder.
B) People often blame you for things you didn't do. (a)
15. A) You are not attracted to prudish people.
B) You are not attracted to unkempt people. (r)
16. A) You want badly to "belong." (D)
B) You don't care whether you "belong" or not.
17. A) You like a clean, neat house. (R)
B) You like good food.
18. A) You can never forget that love is more than just sex. (R)
B) You can take pleasure in sex as sex.
19. A) You are always on the lookout for new ways of attacking a problem.
B) In general, you find the tried-and-true methods work best. (r)
20. A) You are rebellious.
B) You like discipline. (D)
21. A) You don't like to gamble on getting a good break. (a)
B) You usually figure on getting a good break.
22. A) You get more credit than you deserve.
B) You get less credit than you deserve. (a)
23. A) You get into scraps you didn't start. (a)
B) When you get into trouble it is almost always your fault.
24. A) Most everybody lets you know directly what they think of you.
B) Some people are secretly trying to get the better of you. (h)
25. A) You positively like to be different from your immediate associates.
B) Being different from your immediate associates makes you uncomfortable. (D)
26. A) People are either your friends or your enemies. (r)
B) People are rarely either real friends or real enemies.
27. A) Your hardest battles are with other people rather than with yourself. (h)
B) You are cocky.
28. A) You could like anyone if you tried.

- B) There are some people you know you could never like. (H)
29. A) You are forgetful.
B) You have a meticulous memory. (R)
30. A) There are some people you would like to tell off. (h)
B) You are occasionally taken in.
31. A) People criticize you unjustly. (a)
B) People give you more breaks than you deserve.
32. A) You are charming.
B) You are firm and resolute. (R)
33. A) Disappointments affect you so little that you seldom think about them twice.
B) Your daydreams are often about things that can never come true. (A)
34. A) You would like to counsel a friend on his personal problem.
B) You would like to give first aid to a friend. (d)
35. A) You collect things. (R)
B) You lose things.
36. A) You like haphazard living.
B) You like routine. (R)
37. A) Stuffed-shirts amuse you.
B) Stuffed-shirts get under your skin. (h)
38. A) You keep calm in an emergency.
B) You can obey orders. (d)
39. A) You are difficult to please.
B) You like to do favors. (D)
40. A) You are aware of dripping water in the kitchen. (r)
B) You are not observant.
41. A) You don't mind a coward.
B) You can't stand a coward. (h)
42. A) You just can't stay mad even when you think you should.
B) There are some people you would like to take apart. (H)
43. A) You admire spontaneity in people.
B) You admire efficiency in people. (R)
44. A) You don't particularly like to march.
B) You like to march with a group you feel proud to belong to. (d)
45. A) You need someone in whom you can confide completely. (D)
B) You are selfish.
46. A) You play fair. (D)
B) You are an individualist.
47. A) There are some magazines to which you particularly turn for the substantiation of your political ideas. (d)
B) Your political ideas tend to be peculiar to yourself.
48. A) You can't help feeling antagonistic to people who hold important opinions radically different from yours. (h)
B) You like a lot of people who disagree with you violently on important issues.
49. A) Your interest in general principles occasionally gets you up in the clouds.
B) You are a stickler for precision. (R)
50. A) You have felt so sorry for someone you have cried.
B) You have gotten so mad you cried. (h)
51. A) Yours is a quick and ready sympathy.
B) You are stern. (H)
52. A) You are independent.
B) You are loyal. (D)
53. A) You are talkative.
B) Often you're sure you've forgotten something important. (A)
54. A) You would be happier if you felt more secure. (a)
B) You would be happier if you were less gullible.
55. A) You never change your basic beliefs. (r)
B) All your beliefs are open to debate.
56. A) You follow your conscience.
B) You have ethical standards which you follow. (d)
57. A) You are very proud of your membership in some groups. (d)
B) You don't go for groups.
58. A) You are indifferent to most people.
B) You like or you dislike people. (r)
59. A) You don't worry about physical disorders.
B) Sometimes you figure you're a sure thing for ulcers. (a)
60. A) You are dogmatic. (R)
B) You are sloppy.
61. A) There are some people you admire so much you would not question their opinion. (D)
B) You don't admire anybody very much.
62. A) Concerning your past actions you figure, "If I did it, it can't be too bad."
B) If you had your life to live over, there would be a lot of things you'd do differently. (A)
63. A) You admire careful, rigorous thinking. (r)
B) You admire brilliant, penetrating thinking.
64. A) The details of life are important to you. (R)
B) You are often thoughtless.
65. A) You are well coordinated. (r)
B) You seek new opinions.
66. A) You are self-confident.
B) You are a good Joe. (A)
67. A) You are often depressed for no good reason. (A)
B) You always feel life is very much worth living.
68. A) You would like to work by yourself.
B) You would like to work in a good organization. (D)
69. A) You have often felt sick when you remember how badly you have treated someone.
B) You don't worry about people you hurt, when they deserve it. (H)
70. A) You could use some good counsel. (D)
B) You could do with a good friend.
71. A) You are eccentric.
B) You could easily identify yourself with a great cause. (D)
72. A) You like variety.
B) You always finish a job you start. (R)
73. A) You can be squeamish.
B) You can be brutal. (H)
74. A) You are thick skinned.
B) You are easily slighted. (A)
75. A) You would be more successful if you were less optimistic.
B) You would be more successful if you had more confidence. (A)
76. A) You are a fault-finder. (H)
B) You often have the blues.
77. A) You could see yourself as a loyal follower of a great man. (D)
B) You are a mediocre team-man.
78. A) You are touchy about some things. (A)
B) You are not easily insulted.
79. A) Being yourself means being different.
B) Being yourself has nothing to do with being different. (D)
80. A) You are orderly. (R)
B) You are cheerful.
81. A) You are sensible. (R)
B) You are spontaneous.
82. A) You feel no necessity to like people. (H)
B) Disliking someone makes you as uncomfortable as someone's disliking you.
83. A) People can get very nearly anything from you if they cry.
B) You can't stand "dramatics." (H)
84. A) When you hurt someone inadvertently, you feel no guilt. (H)
B) You feel almost as badly when you hurt some-

- one inadvertently as when you meant to hurt them.
85. A) You are methodical. (R)
B) You are curious.
86. A) Nobody puts anything over on you. (H)
B) You get over a humiliating experience quickly.
87. A) You are a delightful person.
B) You are an efficient person. (R)
88. A) People often reject your suggestions not on their merits but because it was you who made them. (A)
B) People as a rule give you credit for more wisdom than you possess.
89. A) You would describe your thinking as "original."
B) You would describe your thinking as "accurate." (R)
90. A) Your ambition is mostly personal.
B) Your ambition is not so much for yourself as for certain ideals which you hold. (D)
91. A) At times you think you are no good at all. (A)
B) You never seriously doubt your own abilities.
92. A) You tend to do a lot of "soul-searching." (A)
B) In general you concentrate pretty much on the problems in front of you.
93. A) You can't stand people who are politically and socially ignorant. (H)
B) You don't place much emphasis upon how ignorant a man is in politics and social affairs.
94. A) There are people who are an inspiration to you. (D)
B) Mostly you are inspired by ideas.
95. A) You never say anything nasty just to be mean.
B) Lots of times you just can't help saying what a dirty so-and-so you think someone is. (H)
96. A) You notice when window shades are not drawn. (R)
B) You are often neglectful.
97. A) Reciting to yourself your own name has a way of calling out the best in you.
B) Those things which symbolize the groups to which you belong call out the best in you. (D)
98. A) You try to be creative in your approach to things.
B) You try to be sound in your approach to things. (R)
99. A) You don't mind apologizing as long as you're even partly wrong.
B) You find it hard to apologize, particularly when the other fellow is as much to blame as you are. (H)
100. A) Even when you are with your friends you feel lonely much of the time. (A)
B) Loneliness for you is just a feeling, which doesn't last long.
101. A) Nervousness keeps you from being as popular as you should be. (A)
B) You are very popular as you are.
102. A) You try to keep a tight rein on yourself at all times. (R)
B) Much of the time you rush things without thinking.
103. A) You like parties where people just get together and let things develop.
B) You like well-arranged parties. (R)
104. A) In many things you are fickle, jumping from one thing to another.
B) You always stick to a plan once you have decided upon it. (R)
105. A) Lots of times people "on the outside" see your problems more clearly than you do.
B) If there is anything which gets you, it is a guy who thinks he knows your problems better than you do. (H)
106. A) You are upset when a thing is not done the right way. (R)
B) A great deal of the time you aren't sure what the right way would be.
107. A) You feel sympathetic toward people who tend to hang on to their griefs and troubles.
B) You have no sympathy for people who aggravate their own problems. (H)
108. A) You can't forget certain failures and humiliations in your past. (A)
B) As you look back on your life you can see things have been for the good.
109. A) You can feel, not just intellectually accept, the power of the people. (D)
B) The only power which you really feel is the power within yourself.
110. A) You prefer work which requires close attention to details. (R)
B) You'd go nuts doing precision work.
111. A) You are responsible to other people, those you love, those who depend on you, and so on. (D)
B) You are responsible to yourself, your own ideals, ambitions, and so on.
112. A) Often times you discover that you've left a door open, a cigarette lighted, a letter unmailed, and so on.
B) You check and recheck things to make sure you have taken care of them. (R)
113. A) You have often had trouble sleeping. (A)
B) You have always been a sound sleeper.
114. A) You like the aloneness and independence of a bustling city street.
B) You like the atmosphere of a crowd of people who meet for a common purpose. (D)
115. A) There are days when you can't make even small decisions. (A)
B) You sometimes mess up a small thing but you rarely go wrong on the big ones.
116. A) You like a person to be modest even if he is good. (D)
B) You like a person to say he is good, provided he is.
117. A) You feel that a person should dedicate himself to something. (D)
B) You are inclined to be self-centered.
118. A) Your first impressions of people are usually correct. (H)
B) You find that as you get to know a person, you get to like him.
119. A) You have solved most of your personal problems.
B) You seem to have a lot of unsolved personal problems. (A)
120. A) To see any other person suffer would upset you very much.
B) To inflict pain on some people would not be altogether unpleasant to you. (H)
121. A) You wake up fresh and rested almost every morning.
B) There are many days you dread going through with. (A)
122. A) You used to keep a diary. (R)
B) Oftentimes you forget important dates in your life, like the day you graduated from high school.
123. A) Your emotions are not easily aroused.
B) You are easily excited or depressed. (A)
124. A) You feel "at sea" much of the time. (A)
B) Psychologically you are always in good shape.
125. A) You often make fun of other people. (H)
B) You often make fun of yourself.
126. A) You are suspicious. (H)
B) You are trusting to a fault.
127. A) You are easy-come easy-go about material things.

- B) You try to conserve and plan for the future. (R)
128. A) You live for the general good first. (D)
B) You live for yourself first.
129. A) You have had periods when you lost sleep worrying. (A)
B) You simply aren't a worrier.
130. A) You often ask people for advice. (D)
B) People often imply that you think you are better than you are.
131. A) You forgive people easily.
B) When people apologize, you can't help feeling: actions speak louder than words. (H)
132. A) You like to feel a member of a good, solid group. (D)
B) For some reason, you are a bit allergic to groups.
133. A) You think of yourself as first of all an individual person.
B) You think of yourself as first of all a social being responsible to society and those who think like you. (D)
134. A) Sometimes you change your way of talking and the things you say in order to win popularity in a group. (D)
B) You've yet to find a group of people who were worth much effort to get in with.
135. A) If someone hurts you, you try to get even with him. (H)
B) You are even-tempered.
136. A) Your emotions rarely get in your way.
B) You have a lot of "pent-up" feelings. (A)
137. A) You feel that your personal life is too serious to be joked about. (A)
B) There is practically nothing you mind being kidded about.
138. A) You like to gamble, in life as well as games.
B) You feel your best chance of success is "to play your cards close to your chest." (R)
139. A) You are self-reliant.
B) You can be depended on. (D)
140. A) One of the nicer things about vacations is that you don't have to be well-groomed all the time.
B) You don't like to visit in a poorly kept house. (R)
141. A) There are a lot of days when no matter how hard you work you just don't seem to accomplish anything. (A)
B) You always manage to get something done.
142. A) You have confidence in yourself.
B) You have confidence in things you and those like you stand for. (D)
143. A) You tend to be forgetful in money matters.
B) Under any circumstance you would keep a close watch on your financial affairs. (R)
144. A) The truth of the matter is that you haven't "found yourself." (A)
B) You feel altogether adequate to carrying out the goals you have set for yourself.
145. A) People talk flatteringly of you when you are not around.
B) People talk critically of you when you are not around. (A)
146. A) You admire the efficiency of a well-run organization. (R)
B) You admire the drive of a well-run organization.
147. A) You don't care much for "lone wolves." (D)
B) You are a "lone-wolf" yourself.
148. A) Some of your friends think that your ideas are impractical, if not a bit wild.
B) You stick to matter of fact as much as possible. (R)
149. A) There are a lot of people you know who don't like you very much.
B) You lack self-confidence. (A)
150. A) You are often nasty. (H)
B) You are very considerate.
151. A) You always work toward tangible and clearly defined results. (R)
B) Much of the time you aren't precisely sure what you're working towards.
152. A) Physically, you are "in the pink" almost all of the time.
B) You have trouble concentrating. (A)
153. A) You enjoy confused and noisy but basically friendly debate.
B) You like a smooth-running organization. (D)
154. A) You like variety.
B) You like order. (R)
155. A) You are forthright. (R)
B) You are carefree.
156. A) You feel rather sorry for the habitual complainer.
B) Whining, complaining people infuriate you. (H)
157. A) Once you resolve to do a thing you always carry through. (R)
B) Concerning your own resolutions you figure: I made it, I can break it.
158. A) Disliking people makes you feel uncomfortable.
B) There are some people that you dislike so much that you are pleased when they get what is coming to them. (H)
159. A) You are finicky about your personal habits and dress. (R)
B) You tend to neglect your personal appearance.
160. A) You like individual sports.
B) You like team sports. (D)
161. A) You tended to be solitary as a child.
B) When you were a child, you belonged to a gang that stuck together through thick and thin. (D)
162. A) You admire decisive, forthright action for its own sake. (D)
B) Sometimes you like to be alone.
163. A) You don't belittle people.
B) You hate people who are condescending to you. (H)
164. A) You are a slob.
B) You are a snob. (H)
165. A) Sometimes you feel it is necessary to hurt someone for his own good. (H)
B) You would have a hard time hurting someone even if you thought it might do some good.
166. A) You are a perfectionist. (R)
B) You like "big" ideas.
167. A) You don't think much about your past.
B) There are some things in your past you can't seem to get out of your mind. (A)
168. A) You often find that people who are antagonistic to you would really like to be friendly.
B) You often notice, beneath a person's surface friendliness, a deeper hostility. (H)
169. A) In most things you have "a golden touch."
B) You have days, even weeks, when everything you do seems to go wrong. (A)
170. A) You hate to force your opinion on anyone.
B) Lots of times you just can't argue with people—you have to tell them what's what. (H)
171. A) You are harsh in your judgment of people. (H)
B) You are lenient in your judgment of people.
172. A) You take jibes and insults in your stride.
B) You are more "sensitive" than most people. (A)
173. A) You sometimes wonder what all the struggle of life is about anyway. (A)
B) You don't worry much about "the struggle of life" but stick to the problems at hand.
174. A) You are almost always alert. (R)
B) Sometimes you just don't pay any attention to what's going on.

175. A) You enjoy the excitement of a crowd. (D)
B) You rather dislike crowds.
176. A) Generally you let a dirty crack go without retaliating.
B) You don't believe in "turning the other cheek." (H)
177. A) In group activities you get more than your share of attention.
B) In group activities you are sometimes glossed over. (A)
178. A) You can be "hard." (H)
B) You are a sucker for a good "hard luck story."
179. A) Sometimes you do things just "to show" somebody. (H)
B) You make a point of not throwing a fellow's mistakes up to him.
180. A) You have made more progress in life than the great majority of people your age.
B) Sometimes you feel as if life is passing you by. (A)
181. A) You enjoy social gatherings just to be with people. (D)
B) You don't enjoy big parties.
182. A) You want to be judged solely on your own merits.
B) You don't mind being judged by your affiliations. (D)
183. A) People ignore you. (A)
B) People "put you on a pedestal."
184. A) You find it difficult to break with familiar and pleasant surroundings. (A)
B) You feel at home almost anywhere.
185. A) You have often been criticized as not sufficiently humble.
B) You place great emphasis upon being a true friend. (D)
186. A) You live from day to day.
B) You try to plan your future in great detail. (R)
187. A) When you were a child you didn't care to be a member of a crowd or gang.
B) You have always been a "joiner." (D)
188. A) People often resent your giving them constructive criticism. (H)
B) You don't give advice unless the other guy asks you to, and not always then.
189. A) You enjoy parades but you have no desire to be in them.
B) When you see a parade go by, you feel like going out and marching with them. (D)
190. A) You are pretty self-satisfied.
B) You long to be at peace with yourself. (A)
191. A) You are exceptionally healthy.
B) You are moody. (A)
192. A) You often try to "feel" your way into someone else's point of view.
B) You treat people according to what they deserve. (H)
193. A) There are people in the world for whom you feel nothing but hatred. (H)
B) You don't hate anybody.
194. A) You tend to be absent-minded.
B) You are thorough. (R)
195. A) Sometimes you say hard things about people. (H)
B) You bend over backwards not to talk anyone down.
196. A) You often find yourself questioning the loyalty of your friends. (A)
B) The loyalty of your friends is not a matter of great moment to you.
197. A) People sometimes criticize you for being "soft."
B) It does not bother you particularly to see animals suffer. (H)

Psychological Monographs: General and Applied

Effects on Clients of a Reflective and a Leading
Type of Psychotherapy¹

JEFFERSON D. ASHBY, DONALD H. FORD, BERNARD G. GUERNEY, JR.,

AND LOUISE F. GUERNEY

With an Introduction by

WILLIAM U. SNYDER

Pennsylvania State University

INTRODUCTION

THE FOUR integrated studies described in the pages following this introduction are interesting in several respects. *First*, the design was unique. The researchers have compared a relatively nondirective method of treatment with a somewhat interpretive method, in order to determine whether one approach induces more resistance in the client than the other, whether one produces more dependency, and whether clients with certain personality characteristics relate better to the therapist in one treatment-approach than the other. In order to make this comparison, training in the use of both treatment methods was given to 10 therapists. Then, during a time interval of one semester, almost all of the clients asking for help at the Psychological Clinic of the Pennsylvania State University were *randomly* assigned to these 10 therapists and to the two treatment conditions. Each therapist worked with four clients, using a nondirective approach with two and a more

interpretive approach with two. Analyses of variance were completed on the data from the two samples. This design made possible a comparison of the two therapeutic approaches, a comparison of the therapists with each other, and an evaluation of the effects of a particular therapist using a particular therapeutic approach. Contributing to the strength of the design was the fact that the experimenters themselves did not act as therapists.

A *second* aspect of these studies that is of interest is that the investigators found it quite possible to classify client and therapist verbal behavior directly from the recordings, rather than from typed transcriptions of the interviews. In fact, they felt that their classifications were probably more accurate under these circumstances, since voice inflections were often crucial in determining how to code a particular response.

A *third* aspect of interest is some rather surprising reactions by the therapists. For example, in spite of the fact that all 10 therapists seemed quite willing to participate in the study, and the experimenters used considerable tact in working with them, some of the therapists did not observe the limits of the study. Four of them decided that "they

¹This report constitutes an integration and condensation of four interrelated doctoral dissertations (1, 9, 16, 17) from the Pennsylvania State University. The studies were conducted jointly under the supervision of William U. Snyder, Leon Gorlow, and Alec J. Slivinske. William Ray served as statistical consultant for the studies.

knew best" and consequently deliberately used a different technique if they felt that it was in the best interests of the client. Although several explanations are possible, the one that seems most likely to the experimenters, in view of the information available, is that their personal needs were too strong for them to be willing to remain within the limits of the experiment. Three of the four were rather authoritarian in their relationship with their clients, and the fourth had a very strong need to be non-threatening. Another thing that wasn't anticipated is that, although the therapists could learn quite satisfactorily how to play the roles required in each treatment method, this intellectual perception of what was expected of them did not necessarily take precedence over their personal needs to do what they thought best. A sociometric measure answered by fellow therapists, and used in the hope that it would provide an index of the stimulus value the therapists held for their clients, proved not to be of value. The attempt to discover positive correlations between the therapists' personality characteristics and clients' behavior was unsuccessful. Nothing is as indicative of what a therapist will do in therapy as a recording of his actual in-therapy behavior! Another finding of interest concerning therapists is that clients seem able to relate satisfactorily not only to friendly, non-threatening therapists, but also to authoritarian therapists who engender confidence. This matter of what therapist personality factors are conducive to an effective therapeutic relationship continues to be a challenging one. The writer has observed that, contrary to his former suppositions, it is possible for a therapist-in-training who has a number of personal problems to establish an adequate

therapeutic relationship with some clients.

The writer would like to mention very briefly some of the results of the four integrated research studies. (a) The therapists consistently rated their clients as having improved more under an interpretive treatment than under the non-directive one. No other change variables reflected differences between treatments. The leading treatment produced more guarded verbal behavior. Since most therapists expressed a preference for an interpretive type of therapy, the researchers point out therapists' ratings may have been influenced by their therapy preferences. (b) Client pre-therapy personality characteristics seemed to be more important in an interpretive therapy than in a nondirective one in the way they related to defensiveness during the early interviews of therapy. (c) The therapists differed in the amount of "guardedness" and/or "defensiveness" they engendered in their clients. Also, a particular therapist using a particular treatment method produced different effects in his clients in defensiveness. This suggests that for the beginning therapist a subtle interaction of personality characteristics and treatment methods is very important to the amount of defensiveness he produces in his clients. One might say we have known this all along. However, it is of value to have experimental evidence of this fact. We need to obtain more specific information about this tantalizing result. (d) The researchers observed that "openness" and "guardedness" in the client are interrelated and that the one type of response cannot be considered without reference to the other. In other words, when clients seemed really to be involved in therapy, they would state a problem, discuss it, and then, because

it is painful to face oneself and modify one's self concept, they would become somewhat defensive or guarded. This sort of behavior seemed to occur in cyclical form. As a consequence, the researchers were led to the conclusion that perhaps resistance, or defensiveness, may be a "good" sign, in that it indicates that the client is really working on his problems. If he never gets into any real problems he has little need to be defensive. Of course, the "best" therapists would be those who would be able to keep this client defensiveness at a minimum and would not add to it through their own inept behavior.

Illustrative of how puzzling some results can be are the following findings. From some points of view it appeared superficially that the nondirective method produced more desirable results than the interpretive. For example, in the nondirective approach, there were larger percentages of client "openness" and "covert resistance" while the interpretive method had larger percentages of "dependency," "guardedness," and "overt resistance." Further, analysis of the "covert resistance" responses revealed that in the nondirective method 42% of these responses were so classified because the client had made long pauses, while in the interpretive method only 13% of "covert resistance" responses were due to long pauses. Also there were less "blocking" and "interrupting" in the nondirective than in the interpretive therapy. However, from another point of view the more interpretive therapy seemed superior. For example, clients in the interpretive therapy tended to become more positive in their feelings toward therapy, as measured by a rating scale completed at the end of the fourth and again at the end of the eighth interview, whereas clients in the more non-

directive therapy tended to become more negative or defensive. Also, therapists were able to hold clients in therapy better in the interpretive situation than in the nondirective one. Both therapies seemed to have certain aspects which produce favorable reactions in some clients.

Another puzzling result was in regard to the relationship between therapist personality and client "guardedness." If "guardedness" scores were considered independently, the lowest scores in this category were with therapists who were friendly, uncritical, and took a conversational approach, rather than focusing on client problems. Also, clients of "friendly" therapists tended to show more decrease in maladjustment scores, while clients of warm, accepting, but dynamically sensitive therapists tended to grow worse in adjustment, although this change was not statistically significant. It would appear, then, that the therapist who is friendly and uncritical and who does not focus as consistently on problems produces the least client "guardedness" and the greatest decrease in "maladjustment," while the therapist who is warm, accepting, but also aware of the client's problems and his motivation, produces more "guardedness" and more "maladjustment." The researchers in these studies chose to believe that the "guardedness" was a necessary concomitant of facing unpleasant facts about oneself, and that the seeming increase in maladjustment was a temporary one which would be replaced by a decrease in maladjustment as therapy progressed.

Lest the present studies should tempt the reader to make unwarranted generalizations, the writer would like to mention some limitations of the studies. First, although the two therapies were very different in some respects they were

similar in others. The interpretive therapy was not extremely so, and the non-directive therapy was not completely "pure." Second, the treatment lasted only through one semester, with an average of about thirteen interviews. Third, the therapists were only moderately experienced. Fourth, the attitudes of the therapists were such that they appeared to have more confidence in the interpretive therapy than in the non-directive therapy. This lack of confidence could have influenced the effectiveness of the therapies in subtle ways.

The results of these four studies make us more aware than ever of the need for continued research on the nature of the relationship between the therapist and the client. We need to be able to identify the personality characteristics which enable a therapist to function in a maximally effective manner, and to delineate those techniques which will produce the most rapid and least painful progress for the client. It appears important to explore client and therapist interview behavior in relationship to therapeutic outcomes. In the writer's opinion such research will be more likely to be useful if it is based, at least in part, upon what actually happens in the therapeutic interview.

NEED AND STATEMENT OF THE PROBLEM

Over the span of years since the introduction of Rogers' *Counseling and Psychotherapy* (23), interest and research in the area of psychotherapy have constantly expanded (2, 3, 8, 10, 21, 25, 29, 31). However, the need for continued research and the development of more systematic theory is evident from our too limited knowledge of the therapeutic process.

Recognizing the need for research on

psychotherapy, and particularly the need for more comprehensive and better designed research, the writers set out to develop a research project which incorporated a formal experimental design, which encompassed numerous variables related to the therapeutic process, and which involved more adequate samples of clients and therapists. The fundamental purpose of the project was to analyze many different variables and to assess their relationship to therapy. The list of problems investigated in this study follows.

Effects of Leading and Reflective Therapy

1. *Do pretherapy characteristics of clients relate differentially to the clients' reactions to a reflective and to a leading type of therapy?* An answer was sought in relation to each of the following client pretherapy characteristic variables: (a) need for autonomy, (b) need for succorance, (c) need for deference, (d) need for aggression, (e) tolerance-intolerance of cognitive ambiguity, and (f) defensiveness. The relationship of each of the preceding variables to client reactions to leading and to reflective types of therapy was explored with respect to (a) the therapeutic relationship as viewed by clients, and (b) the amount of defensive verbal behavior exhibited by clients in therapeutic interviews.

2. *Does clients' verbal behavior in therapy differ in a reflective and a leading type of therapy?* The client verbal behavior variables explored were (a) dependence, (b) openness, (c) guardedness, (d) covert resistance, and (e) overt resistance.

3. *Does the relationship between a client and his therapist differ in a reflective and a leading type of therapy?* The client

relationship variables consisted of clients' subjective positive and defensive reactions to therapy and therapist. The therapist relationship variables consisted of the therapists' subjective positive and negative reactions to the clients and the therapy situation. An answer was sought at the fourth and eighth interviews.

4. *Do changes in clients through therapy differ in a reflective and a leading type of therapy?* The client change variables investigated were (a) level of maladjustment, (b) anxiety, (c) defensiveness, (d) dependency, (e) positive attitudes toward self, (f) positive attitudes toward others, and (g) therapists' evaluation of client changes.

Effects of Therapists as Individuals

1. *Are personal characteristics of therapists related to the effects therapists have on their clients?* The therapist personal characteristic variables investigated were (a) ability to enter the phenomenological field of another, (b) sympathetic interest, (c) acceptance of others, (d) social stimulus value, (e) need to aggrandize the self, and (f) aggression. The effects on clients which these characteristics might have were explored with respect to (a) clients' defensive verbal behavior in therapy interviews, (b) the client relationship variables, and (c) client changes in maladjustment through therapy.

2. *Are there differences among therapists in the way they affect clients' verbal behavior in therapy?* The client verbal behavior variables explored were (a) dependence, (b) openness, (c) guardedness, (d) covert resistance, and (e) overt resistance.

3. *Are there differences among therapists in the relationship they establish with their clients?* This question was ex-

plored with respect to clients' positive and defensive views of the relationship and to therapists' positive and negative views of the relationship at both the fourth and eighth interviews.

4. *Are there differences among therapists in the changes they produce in their clients during therapy?* The client change variables investigated were (a) level of maladjustment, (b) anxiety, (c) defensiveness, (d) dependency, (e) positive attitudes toward self, (f) positive attitudes toward others, and (g) therapists' evaluation of client changes.

Effects of the Interaction Between Therapists and Type of Therapy Administered

1. *Is client verbal behavior in therapy affected by the interaction of the therapist as an individual with the type of therapy he is employing?* The client verbal behavior variables explored were (a) dependence, (b) openness, (c) guardedness, (d) covert resistance, and (e) overt resistance.

2. *Is the therapeutic relationship affected by the interaction of therapists as individuals with the type of therapy being employed?* This question was explored with respect to clients' positive and defensive views of the relationship and to therapists' positive and negative views of the relationship at both the fourth and eighth interviews.

3. *Is the extent of change in clients through therapy affected by the interaction of therapists as individuals with the type of therapy they are employing?* The client change variables investigated were (a) level of maladjustment, (b) anxiety, (c) defensiveness, (d) dependency, (e) positive attitudes toward self, (f) positive attitudes toward others, (g) therapists' evaluation of change scores.

EXPERIMENTAL DESIGN AND PROCEDURES

Independent Variables

The effects of two independent variables were examined in this study. The first was the type of therapy administered and the second was the therapist as an individual. The type of therapy was manipulated by defining two families of therapist verbal responses.

Reflective Therapy

This family of responses included restatement of content, reflection of feeling, nondirective leads, and nondirective structuring responses. This therapy was built largely on the Rogerian approach (23, 24). Therapists' behavior was guided by the following working hypotheses.

The therapist attempts to create a warm, acceptant, understanding, noncritical psychological atmosphere; to understand and accept the feelings which the client experiences as a result of his perception; and to communicate this acceptance and understanding to the client.

The therapist believes the client has within himself a capacity to understand himself and a capacity and tendency to reorganize himself. The therapist also believes that, in a warm, acceptant, understanding, and noncritical atmosphere, the client will reorganize himself at a rate and to a depth most appropriate for him.

It is necessary for the therapist to accept and clarify only those thoughts and feelings which the therapist believes are in the client's present phenomenological field. These thoughts and feelings must be strongly implied by the client himself, if they are not explicitly communicated either verbally or nonverbally. By consistently maintaining this role, the therapist enables the client to eliminate his need for defenses in the therapeutic situation, to recognize his conflicts, his emotional reactions and needs, and to bring about a self-reorganization of his patterns of perception and behavior.

Leading Therapy

The second family of responses was composed of directive leads, interpretations, directive structuring, approval, encouragement, suggestion, advice, information giving, and persuasion. The leading therapy was based largely on the approaches of Dollard and Miller (6) and Fromm-

Reichmann (11). It was guided by the following working hypotheses.

The therapist attempts to create a warm, accepting, understanding, noncritical psychological atmosphere; to contrast the client's report of his situation and difficulties with an objective reality as the therapist deduces it; to formulate hypotheses about the defenses which protect the conflicts; and to intervene in such a way that he helps the client understand the nature and function of the defenses. The therapist may then help the client in coping directly with underlying conflicts at a level which the therapist deems advisable and feasible within the limitations of time and the client's personal dynamics. He thus helps the client to become reoriented in terms of reality.

The therapist believes that the client has a capacity to learn new behavior patterns, but that this capacity is not being utilized effectively because the client's defenses, inappropriate reaction patterns, and fears prohibit his becoming aware of, and trying out, alternative patterns of perception and behavior.

It is necessary for the therapist to introduce, or direct attention to, factors not within the client's present awareness, in order to make the client aware of his defenses, to help him modify them or eliminate the need for them, to recognize his conflicts, emotional reactions, and needs, and to bring the client to adopt alternative patterns of perception and behavior.

Therapists

Therapists as individuals constituted the second independent variable. Ten therapists were used in this study. It should be noted that the authors did not participate as therapists. The therapists were all advanced graduate students ranging in age from 24 to 30 years. Six of the therapists had internship experience in a medical setting approximating one year or more. One had experience as a mental hospital attendant, caseworker, interviewer, and college counselor. One had approximately a year's experience with vocational and personal counseling in a university setting. Another had a year of experience with vocational and personal counseling in a university setting plus work in a school for delinquent girls under the supervision of a psychiatrist and a social worker. Another had worked three months doing casework in a settlement house. In addition, some had work experience in military, private, and public settings. They reported that their therapeutic biases were still in a formative stage, though at the time of the experiment all but one reported an inclination toward a leading type of treatment. All had

had supervised experience in vocational and educational counseling in the Pennsylvania State University Psychological Clinic.

Prior to the beginning of the experiment, the 10 therapists had just completed a course in psychotherapy. In the course, they studied systems representative of both types of therapy. In addition, each person had carried two or more therapy cases under the supervision of an experienced clinical psychologist. The therapists also participated in a training program devised especially in preparation for this research. Through readings, practice with typescripts of previous cases, role playing, and discussions, the therapists were familiarized with the response families and given practice in their use. On the basis of the final role-playing session, all of the therapists were judged to be differentiating the two types of therapy.

A system for coding therapists' responses was devised and used to code responses from eight recorded interviews for each therapist. In a pilot study designed to demonstrate the reliability of the coding system, three of four judges coding independently agreed 92% of the time as to which of the response families a given response belonged. Three of four judges agreed 82% of the time that a given response belonged in one of 15 different categories. The coding system used was patterned after that of a previous research project (30). A criterion of approximately two-thirds of each therapist's experimental responses in the appropriate response family for all clients was established as the minimum acceptable differentiation of treatments. In addition, for a therapist to qualify, at least 60% of all his experimental responses had to be in the appropriate response family for each individual client. Six of the 10 therapists met the criteria. *Only these six therapists and their 24 clients are included in the principal statistical analyses of this study.* Some of the qualitative observations made in the study rest on all 10 therapists and their 40 clients. Table 1 shows the distribution of responses in the 15 response categories for all 10 therapists. Clearly, restatement of content and clarification of feeling responses were emphasized in the reflective therapy, while interpretations and directive leads were emphasized in the leading therapy. If the nonexperimental categories of XESCFD, XTR, XUN, and XUNT (see Table 1) are excluded and the responses of the six therapists meeting the criteria are examined, 89% of their reflective therapy responses were appropriate to that therapy, while 81.5% of their leading therapy responses were appropriate to that therapy. The nondirective emphasis compares favorably with previous research describing the verbal pattern of Rog-

TABLE 1
THE NUMBER AND PROPORTION OF THERAPISTS' RESPONSES IN EACH RESPONSE CATEGORY FOR EACH OF THE EXPERIMENTAL THERAPIES^a

Response Categories ^b	Reflective Therapy		Leading Therapy	
	Number of Responses	% of Total Responses	Number of Responses	% of Total Responses
XCS	18	1.1	1	.1
XND	16	1.0	14	.7
XRC	547	34.4	168	8.6
XCF	509	32.1	123	6.3
XESCFD	35	2.2	54	2.8
XTR	30	1.9	52	2.7
XUN	36	2.3	36	1.8
XUNT	51	3.2	59	3.0
XDS	6	.4	32	1.6
XDL	155	9.8	880	45.1
XIT	130	8.2	358	18.3
XIF	39	2.4	90	5.1
XAER	11	.7	44	2.2
XSAP	3	.2	32	1.6
XDC	1	.1	2	.1

^a Based on part or all of 10 therapists' responses on 40 records for each treatment.

^b XCS—Nondirective structuring.

XND—Nondirective leads.

XRC—Restatement of content.

XCF—Clarification of feeling.

XESCFD—Ending contact, series, or free discussion.

XTR—Therapist reaction.

XUN—Unclassifiable.

XUNT—Unclassifiable, recording unclear.

XDS—Directive structuring.

XDL—Directive leads.

XIT—Interpretation.

XIF—Information giving.

XAER—Approval, encouragement, reassurance.

XDC—Direct criticism.

XSAP—Suggestion, advice, persuasion.

erian therapy (26, 29, 30). Research on leading treatment response-patterns is not available for comparison.

Dependent Variables

Client Variables

The major criteria used in selecting the client variables were: (a) the measure had to have a logical and meaningful relationship to the therapeutic process; (b) the measure had to be obtainable without placing unreasonable demands

on the client population; and (c) the measure should reasonably be expected to demonstrate differences between a reflective and a leading type of psychotherapy.

Client verbal behavior in therapy. Five main variables and one composite variable were based on clients' verbal behavior in the first four interviews. A tentative classification system based in part on previous research (4, 5, 15, 18, 19, 27) was devised by the four authors. A pilot study on nonexperimental recorded interviews produced several modifications of the tentative system. The coding rules were intended to identify responses reflecting general sets, rather than responses specific to individual therapists' statements. Descriptions of the verbal behavior summarizing the client verbal behavior variables follow.

1. *Dependence:* The extent to which the client asks the therapist for his opinions, advice, information, evaluation, and instruction, or demonstrates a need for structuring from the therapist.

2. *Openness:* The extent to which the client freely discusses his problems, his deviations from the "normal," his culturally frowned-upon traits, behavior, and motivations; and, in general, his willingness to expose himself to potential criticism and change; especially his willingness to discuss thoroughly those areas which seem most threatening. The client does this without at the same time qualifying, hedging, and engaging in defensive verbal maneuvers.

3. *Guardedness:* The extent to which the client exhibits wariness and hedging in regard to presenting and working on his problems, admitting faults, and exposing himself to potential criticism and change. This includes self-stimulated denial or minimization of his problems or his deviations from the "normal," and denial of culturally undesirable feelings, traits, and motivations. It also includes the need to justify himself or his actions to the therapist, and expectations of criticism from the therapist.

4. *Covert Resistance:* The extent to which the client manifests indirect or impersonalized criticism of the therapist or therapy. It includes blocking, delaying tactics, failure to recall or report things, changing the subject, and interrupting the therapist. It is resistance or hostility toward therapy, therapist, progress in therapy, or toward things which are thought of as being conducive to such progress. But the resistance is not directly expressed verbally; instead, other subtle escapes or hostilities are resorted to by the client.

5. *Overt Resistance:* The extent to which the

client verbalizes criticism—in an open way—of the therapist or the therapeutic method. It includes personal and verbalized opposition to staying within the limits set by the particular kind of therapy which the client is receiving. This is verbalized unwillingness as opposed to "inability" or failure per se.

6. *Defensiveness:* The sum of guardedness, covert resistance, and overt resistance.

A time interval was chosen as the unit of response so that reliability for each client response could be determined when the coding was done aurally from tapes and discs (all coding was conducted in this fashion). The unit of measurement, or the client "response" to be coded, was a 15-second interval of client verbal behavior. Every second of the interview during which the therapist was not talking was regarded as consisting of the client's verbal behavior. Each client response was coded in one of the experimental categories or in a "none" category. In this manner, all of the client's verbal behavior was classified. A client's score on a given variable was the percentage of all his responses which were so classified. In the opinion of the authors, the aural method of coding contained many advantages over transcripts.

The four authors were the coders in the reliability study. In addition to many hours of previous experience with the categories, they underwent a training program of approximately 15 hours. Each coder simultaneously but independently classified client responses on four nonexperimental recorded interviews which were representative of the two experimental therapies. Of the 725 client responses on the records, 468 were classified by three or all judges as falling outside of the experimental categories. Thirty-five per cent, or 257 responses, were placed under one of the dependent variables by two or more judges. In this combined task of locating and categorizing these experimental responses, at least three out of four coders were in agreement on 68% of the responses; at least half of the coders agreed 95% of the time. When coder agreement among the experimental categories alone is considered, excluding differentiation of an experimental from nonexperimental response from the data, all coders agreed on 81% of the client responses. Three or more of the coders agreed on 89% of the responses. These figures compare favorably with reliability studies reported in other investigations (15) in which client responses were coded from transcripts.

On the records classified for the experiment proper, the average client made 162 responses during the 45-minute interview. Of these, 52 (32%) were responses which fell under one of the dependent variable classifications. Table 2 shows

TABLE 2
PERCENTAGE OF ALL CLIENT RESPONSES FALLING UNDER EACH OF THE
MAJOR CLIENT VERBAL BEHAVIOR VARIABLES

	Dependence	Openness	Guardedness	Covert Resistance	Overt Resistance
Reflective Sample	2.26	4.45	3.92	23.58	.63
Leading Sample	2.67	3.47	6.49	16.00	.72
Combined Sample	2.46	3.96	5.20	19.79	.68

the percentage of all client responses falling under each of the major client verbal behavior variables for the clients under the reflective therapy, the leading therapy, and these samples combined.

To obtain some information on the validity—in the sense of agreement with qualified opinion—and objectivity of the experimental coding, a Process Rating Scale was devised. This scale contained the summary descriptions of the five major client verbal behavior variables. Therapists were asked to rate on a five-point scale the extent to which each of their clients exhibited dependence, openness, guardedness, overt resistance, and covert resistance. In order to maintain naivete the therapists were not given the scale until the close of the experiment. To make their rating more comparable in time to the experimental coding, they were asked to recall the behavior of their clients as it was in the first four interviews and rate accordingly. With 38 degrees of freedom, $\pm .31$ and $\pm .40$ are significant at the .05 and .01 levels respectively. The correlations of therapists' ratings with the scores obtained through coding were as follows: dependence .55; openness —.23; guardedness .03; overt resistance .59; and covert resistance .35. Considering the different nature of the types of measurement and the memory distortion that could have entered the ratings, these correlations seem to speak very well for the inherent power of the operational definitions and the reliability and objectivity of measurement in the experimental coding of at least three of the variables: dependence, overt resistance, and covert resistance.

Relationship: The Client Personal Reaction Questionnaire. This questionnaire (called the CPRQ hereafter), constructed by the authors, is composed of two 40-item scales.

One scale is intended to measure *defensive subjective reactions* to therapist and therapy. It includes items reflecting denial, distortion, withdrawal, justification, rationalization, projection, hostility, evasiveness, blocking of thought, blocking of speech, obscuring or confusing issues, anger, fear, criticism, resentment, and self-deprecation.

The second scale is intended to reflect *positive subjective reactions* to therapist and therapy, including a sense of progress, achievement, or accomplishment; feelings of identification and involvement with therapy and/or the therapist; feelings of "safety" and/or security in the therapy situation; satisfaction of needs for acceptance, understanding, help, approval, respect, encouragement; and feelings of respect, admiration, confidence, and gratitude toward the therapist.

The two scales were constructed by having six advanced graduate students in clinical psychology write items to fit definitions for the two scales. Each of the items thus obtained was given a rating from one to four (poorest to best) by each of the four authors. Those items with the highest average ratings were included in the scale. These final items were further screened to avoid overemphasizing one type of reaction. Items which involved making judgments about therapist or therapy such as, "My therapist is well educated," were excluded. Only statements likely to elicit a subjective personal reaction were included, e.g., "My therapist is a nice guy."

Test-retest correlations were computed from data obtained in the experiment. Such correlations have treatment-effects intervening, but are still worth noting. The defensive CPRQ had a test-retest correlation of .79 ($p < .001$). The positive CPRQ had a test-retest correlation of .52 ($p < .01$). The positive and defensive scales obtained at two points in therapy correlated —.35 ($p < .05$) at the fourth interview and —.31 ($p < .05$) at the eighth interview.

The Edwards Personal Preference Schedule. This test (7) measures 15 personality variables which have their origin in a list of manifest needs presented by H. A. Murray and others. Five needs from this test were used as client pre-therapy characteristics. They were the need for deference, autonomy, succorance, dominance, and aggression.

Tolerance-Intolerance of Cognitive Ambiguity Test. This test by Siegel (27) consists of 20 pictures and 20 unrelated statements. Pictures and statements were both randomly selected by Siegel from different groups of magazines. The

client was instructed to compare the statements with the pictures, and to indicate if he felt that any of the persons pictured made any of the statements. He was to match only those he felt were appropriate. The greater the tendency to associate statements with pictures, the lower was the tolerance for ambiguity.

The Mooney Problem Check List. The measures used from this test (20) were the total number of problems checked and the number of words used by the client in summarizing his problem. These two measures had served as measures of client defensiveness in a previous study (30). The assumption was made that the more restricted a client was in admitting and discussing problems, the higher was the level of defensiveness.

The Minnesota Multiphasic Personality Inventory. Six scales derived from the population of items on this inventory were used in the present study: Maladjustment Index (12), Taylor Anxiety Scale (32), Defensiveness Scale (13), Dependency Scale (22), Positive Attitude Toward Self Scale (14), and Positive Attitude Toward Others Scale (14). These variables were intended to reflect the pre- to post-therapy changes.

The Therapist Posttherapy Rating Scale. This scale, developed by the present research group, was modeled after an earlier scale (33). It consists of 27 items, which reflect changes expected to occur in clients undergoing therapy. Items were selected by judges from a population of 77 items as being those most likely to reflect changes resulting from therapy.

Therapist Variables

The therapist-characteristic variables chosen for study were selected with several criteria in mind: (a) their relatively enduring nature, i.e., seeming unalterableness as a result of specific training; (b) their logical relationship to important aspects of prescribed therapist behavior; (c) their measurability; (d) their being relatively unbiased by the psychological sophistication of the therapists; (e) their objectivity of scoring.

Ability to Enter the Phenomenological Field of Another. This was defined as interest in learning about the internal frames of reference of others and being able to see how others perceive and feel in terms of these internal frames of reference. The Intracception scale from the Edwards Personal Preference Schedule, and a

measure of role-playing ability devised by the authors were used to measure this variable. The latter measure was constructed by asking therapists to play the role of a client they know well, while a cotherapist played the role of a Leading Treatment therapist. The therapists understood this to be part of the training program. Ratings of the realism of the role played were assigned for three separate dimensions of role-playing ability. These dimensions were (a) the content, i.e. what the "client" talked about, (b) the reactions that the "client role-player" showed to the therapist's leads, (c) the affect which the role-player displayed while acting as client. Four judges assigned the ratings independently after practice in learning to make reliable judgments. A scoring guide was prepared defining five points for each dimension. Complete agreement was obtained among the four judges on 65% of the judgments, while 94% of the judgments were either in complete agreement or only one step removed from the consensus.

Sympathetic Interest. This was defined as a kindly interest in the activities and thoughts of others and was measured by therapists' scores on the Edwards Nurturance scale.

Acceptance of Others. This was defined as the willingness and/or ability of the therapist to understand and accept what the client has to say without feeling a need to evaluate, judge, or make criticism either openly, or in his own mind. The authors constructed a test called a Test of Clinical Judgment to measure this variable. It consists of items purported to be statements of beliefs, opinions, and values made by unidentified individuals. These items represent viewpoints deviant from those of this culture in general and those of psychologists in particular. However, the items do not express viewpoints so deviant as to warrant their necessarily being conceived of as pathological. Therapists were asked to classify the statements as being made by an adjusted or maladjusted individual. It was reasoned that those who classified fewer items as maladjusted would be those who were most accepting of the values of others. A pilot study with 30 clinical psychology graduate students supported the idea that the items intended to be ambiguous could be viewed as "adjusted" or "maladjusted." Six clearly pathological and six clearly normal statements were included as filler items to provide a frame of reference and as a kind of validity check. Twenty-eight of the thirty persons taking the test in the pilot group classified the filler items as intended by the authors, indicating that items clearly "adjusted" or "maladjusted" could be so classified.

Social Stimulus Value. This was defined as the favorable effect the individual produces on others with whom he has social contact. A so-

ciometric measure devised by the authors was used for this variable. Therapists were asked to select the two most preferred and two least preferred members of their group in five different social situations which involved confiding threatening criticisms, personal friendship, cooperative work, professional supervision, and personal therapy. Scores were derived reflecting the degree of preference for each of the therapists.

Need for Aggrandizing the Self. This was defined as the need to make oneself important by gaining the attention, admiration, and awe of others. Therapists' scores on the Edwards' Exhibitionism scale were used to measure this need.

Aggressiveness. This was measured by the Aggression scale from the Edwards.

Relationship: The Therapist Personal Reaction Questionnaire. This questionnaire (called the TPRQ hereafter), constructed by the authors, is composed of two scales of 35 items each. One scale is intended to reflect Negative Reactions to therapy and client, and includes items reflecting feelings of hostility, resentment, criticism, superiority toward the client; feelings of doubt, discouragement, uncertainty, and failure in regard to progress and accomplishment with the client in therapy; feelings of anxiety, displeasure, discomfort, boredom in anticipation of or in the interviews; feelings of incompetence, inadequacy, ineffectiveness, lack of understanding, and inability to help both in regard to interview behavior and in the long run; feeling disliked, rejected, ridiculed, and pushed.

The Positive Scale reflects feelings of progress, achievement, and accomplishment with the client in therapy; feelings of identification and involvement with the client; feelings of comfort, pleasure, and anticipation in relationship to the interview hour; feelings of respect, admiration, sympathy, and affection for the client; and gratification of existing needs such as those for approval, respect, and therapeutic competence.

The construction of the TPRQ was identical to that of the Client PRQ previously described. Test-retest correlations were obtained by correlating a score obtained at the fourth interview with a score obtained at the eighth interview. The negative scale had a test-retest correlation of .85 ($p < .001$), while the positive scale had a test-retest correlation of .81 ($p < .001$). The positive and negative scales correlated $-.23$ ($p > .10$) at the fourth interview and $-.18$ ($p > .10$) at the eighth interview.

Clients

Most of the clients used in this study were young adults in their twenties whose symptoms were primarily neurotic

in character. Each therapist had two clients in each treatment. Thus, there were 24 clients for the six therapists reported in this study. Seven of the clients were women and 17 were men. There were two women in the reflective and five in the leading therapy. The presenting symptoms often included some reference to unsatisfactory academic performance, since the sample consisted largely of university students. As therapy progressed, however, it was usually apparent that the academic problem was primarily symptomatic. Problems included inability to get along with peers or parents, feelings of inadequacy, sexual frigidity, homosexual impulses, unsatisfactory marital relationships, and disturbing emotions such as anxiety and depression. The intensity of problems ranged from mild to severe. The experiment was terminated at the end of the spring semester because many of the clients were leaving for the summer. The number of interviews completed at termination ranged from one to 23, with an average of 12.8. The fact that a considerable proportion of the clients continued their interviews after the experiment was terminated indicated that they had not completed their therapy at that point. The sample of clients described includes those who completed *four or more interviews*. (In the part of this study concerned with client verbal behavior in therapy, there was no criterion of four interviews. The first four clients completing one or more interviews for each therapist were included. However, this sample differed from the former by only two clients so that the samples are almost identical.)

Procedures

Clients were randomly assigned to

therapists and to the two therapies. Therapists met their clients in 45- to 60-minute interviews twice a week on non-consecutive days. All interviews were recorded on discs or tapes. Before the initial interview, clients completed the MMPI, Edward's Personal Preference Record, the Mooney Problem Check List, and the Tolerance-Intolerance of Cognitive Ambiguity Test. After the fourth, eighth, fifteenth, and terminal interviews, clients and therapists completed Personal Reaction Questionnaires. After the client terminated therapy or the experiment was terminated, whichever occurred first, the client completed the MMPI again. In addition, several tests were completed by therapists early in the experiment.

Clients were never told they were participating in an experiment. They were informed that recordings and test data were used for research purposes. All procedures were handled as a part of regular clinic routine. It is the experimenters' impression that clients were unaware that two types of therapy were being used and no problems arose concerning the therapies. Therapists knew nothing about the variables or questions being explored until after the experiment was completed. Therapists reported experiencing considerable discomfort at the requirement of administering two different therapies. After the experiment, most felt they had benefited professionally from the experience.

Statistical Design

A double classification analysis of variance design was used in the present experiment. This design made it possible to evaluate differences between therapies, differences among therapists, and differences resulting from interaction effects of the two independent variables on the dependent variables. Correlations between pretherapy measures of clients and the dependent variables were not large enough to warrant analysis of covariance.

In the case of the client verbal behavior variables, a client's score on a

given variable was the proportion of all of his responses which fell under that classification. The proportions were transformed to angles for the analyses of variance (28, p. 316).

Correlation procedures were used to relate client pretherapy personality characteristics to client reactions during therapy. Similar procedures were used to relate clients' reactions in therapy to therapist personality characteristics. A $p < .05$ was used as an acceptable level of significance.

RESULTS AND DISCUSSION

Client Pretherapy Characteristics

Tables 3, 4, and 5 report pairs of correlations between client pretherapy characteristics and three client therapy measures.

These correlations suggest that the more aggressive a client, the more verbal defensiveness will be manifested in the leading therapy. Although the correlation in the reflective therapy does not meet the criterion for statistical significance by itself, it is large enough to suggest the possibility that the more aggressive a client the less verbal defensiveness will be manifested in the reflective therapy (Table 3).

Clients who were less willing to discuss their problems on tests prior to therapy also tend to be more defensive in their verbal behavior in the leading therapy, while clients with less pretherapy defensiveness tended to be less defensive verbally in the leading therapy. The willingness to elaborate about problems on a test seemed not to relate to interview verbal defensiveness in the reflective therapy (Table 3).

Deferent clients reacted with more subjective defensive reactions, and less deferent clients reacted with fewer defensive reactions in the leading therapy. There seemed to be no consistent relationship between the trait of deference and subjective defensive reactions of clients toward a reflective therapy (Table 5).

Autonomous clients tended to react with fewer subjective defensive reactions to leading therapy whereas less autonomous clients reacted with more defensive reactions. There seemed to be no consistent relationship between this trait and clients' reactions to the reflective therapy (Table 5).

It is worth noting that three correla-

TABLE 3
CORRELATIONS BETWEEN CLIENT PRETHERAPY MEASURES AND THE CLIENT
VERBAL BEHAVIOR DEFENSIVENESS SCALE

Client Pretherapy Measures	Leading Therapy	Reflective Therapy	Significance of Difference between r 's ^a	p
TICA ^b	-.36	-.10	.82	> .05
Deference	-.12	.00	.61	> .05
Autonomy	.41	-.01	1.31	> .05
Succorance	.02	-.33	1.05	> .05
Dominance	.14	-.06	.58	> .05
Aggression	.42	-.37	2.45	< .02
Mooney: Number of problems	-.28	-.10	.55	> .05
Mooney: Number of words	-.60 ^c	.11	2.33	< .02

^a Based on Fisher's z .

^b Tolerance-Intolerance of Cognitive Ambiguity.

^c $r = .444$ required to be significant at .05 level.

TABLE 4
CORRELATIONS BETWEEN CLIENT PRETHERAPY MEASURES AND CLIENT POSITIVE PERSONAL
REACTION QUESTIONNAIRE AT THE FOURTH INTERVIEW

Client Pretherapy Measures	Leading Therapy	Reflective Therapy	Significance of Differences Between r 's ^a	p
TICA	.38	-.05	1.31	> .05
Deference	-.17	.21	1.11	> .05
Autonomy	-.04	-.44	1.25	> .05
Succorance	.32	.01	.93	> .05
Dominance	-.24	.00	.96	> .05
Aggression	-.15	-.25	.32	> .05
Mooney: Number of problems	.22	-.27	1.46	> .05
Mooney: Number of words	.17	-.05	.64	> .05

^a Based on Fisher's z .

TABLE 5
CORRELATIONS BETWEEN VARIOUS CLIENT PRETHERAPY MEASURES AND THE CLIENT DEFENSIVE
PERSONAL REACTION QUESTIONNAIRE, FROM THE FOURTH INTERVIEW

Client Pretherapy Measures	Leading Therapy	Reflective Therapy	Significance of Difference Between r 's ^a	p
TICA	.29	-.00	1.14	> .05
Deference	.46 ^b	-.28	2.30	< .05
Autonomy	-.52 ^b	.18	2.22	< .05
Succorance	-.00	.12	.61	> .05
Dominance	-.04	-.24	.58	> .05
Aggression	-.00	.08	.50	> .05
Mooney: Number of problems	-.07	.35	1.28	> .05
Mooney: Number of words	-.22	.23	1.31	> .05

^a Based on Fisher's z .

^b $r = .444$ required to be significant at .05 level.

tions for the leading treatment sample were statistically significant while there were no significant correlations for the reflective treatment sample. This suggests that certain client reactions, at least during the first four interviews, are more predictable in a leading than in a reflec-

Differences Between Therapies

Twenty-one analyses of variance were completed. Table 6 reports these data. Only 2 of the 21 showed a statistically significant difference between the two therapies. The similarity of the effects produced on clients by the two therapies

TABLE 6
F RATIOS AND PROBABILITY STATEMENTS FOR THERAPY EFFECTS ON 21 VARIABLES
ON WHICH ANALYSES OF VARIANCE WERE COMPLETED

Variable	Reflective Mean	Leading Mean	F Ratio	p
Client Verbal Behavior Variables:				
Dependence	2.26	2.67	.26	> .20
Guardedness	3.92	6.49	4.02	< .05 > .01
Openness	4.45	3.47	.06	> .20
Covert Resistance	23.58	16.00	.62	> .20
Overt Resistance	.63	.72	.26	> .20
Defensiveness	22.6	21.0	.05	> .20
Relationship Variables:				
Defensive CPRQ—4th	206.5	213.1	.34	> .20
Defensive CPRQ—8th	214.0	197.3	1.98	< .20 > .10
Positive CPRQ—4th	312.6	333.9	.50	> .20
Positive CPRQ—8th	317.3	342.9	2.95	< .20 > .10
Negative TPRQ—4th	72.8	67.0	.47	> .20
Negative TPRQ—8th	74.8	61.3	1.85	< .20 > .10
Positive TPRQ—4th	101.8	97.3	.12	> .20
Positive TPRQ—8th	99.1	105.1	.28	> .20
Change Variables: ^a				
Maladjustment Index	+0.25	-6.09	.94	> .20
Dependency	+1.90	-2.35	.55	> .20
Defensiveness	+0.50	+0.05	.04	> .20
Positive Attitudes Toward Self	+0.35	+1.75	2.27	< .20 > .10
Positive Attitudes Toward Others	+0.70	+0.53	1.16	> .20
Taylor Anxiety Scale	+0.15	-0.45	.38	> .20
Therapist Posttherapy Rating Scale ^b	67.8	91.8	6.32	< .05 > .01

^a All analyses were made on the difference between pre- and post-therapy scores, except for the Taylor and Posttherapy Rating Scale where posttherapy scores were used. A decrease in score indicates a decrease in the trait measured.

^b A higher score on the Therapist Posttherapy Rating Scale indicates a therapist judgment of greater improvement.

tive therapy. It is interesting to note also that the four significant differences in correlation involved measures of client defensiveness, one a behavioral measure of in-therapy resistance and the other a measure of the client's subjective reactions toward therapy. This suggests that defensiveness in a leading therapy is one characteristic which is predictable from client pretherapy characteristics.

is a prominent finding in this area of the study. On the Therapist Posttherapy Rating Scale, therapists consistently rated clients in the leading therapy as showing more improvement than clients in the reflective therapy. This finding must be accepted with qualification, however, since all but one therapist expressed a preference for a leading type of therapy prior to the experiment. Their

preferences may have affected their ratings. Clients' verbal behavior during the first four interviews was more guarded in the leading than in the reflective therapy. This suggests the leading therapy may have been somewhat more threatening. Guardedness will be considered in detail in the discussion of differences among therapists.

Although the levels of significance were not impressive, the consistency with which trends appeared between the fourth and eighth interview on the relationship measures is worth noting. The CPRQ suggests that clients tend to become more defensive in the reflective therapy and more positive in the leading therapy. Therapists tended to become somewhat more negative in their reactions toward clients in the reflective therapy.

Although statistically significant differences between the therapies were few, some trends in the data suggest that the two therapies may have somewhat different effects. Clients in leading therapy tended to become less defensive while clients in reflective therapy tended to become more defensive in their subjective reactions toward therapy. Leading therapy clients tended to show greater positive change in their attitudes toward themselves. Therapists tended to become more negative in their reactions toward clients in reflective therapy. Therapists were able to hold clients better in the leading therapy than in the reflective therapy, and rated clients in leading therapy as more improved. On the other hand, clients in the reflective therapy were less guarded and tended to exhibit less dependence and overt resistance, and more openness than clients in leading therapy. Covert resistance tended to be greater in the reflective therapy, due

largely to a much greater number of client long pauses. Some other components of covert resistance, namely blocking and interruption of the therapist, were noticeably less frequent in the reflective therapy than in the leading therapy.

The writers had opportunities to observe some clients after they completed therapy. These observations indicated that improvements were made by clients in both types of therapy.

Several possible explanations may account for the fact that there were few statistically significant differences between the therapies. The findings may mean that the leading and reflective therapies do not produce very different results. It is also possible that the two do produce different results but that differences did not show up because of the limited power of the experiment. The measures used may have been inadequate, or it may be that the variables themselves were not appropriate to demonstrate differences which actually exist.

Differences Among Therapists

Statistically significant differences among therapists appeared in 4 of the 21 analyses, as shown in Table 7. Three of the four differences appeared on the relationship variables. Therapists appear to produce differing degrees of subjective defensive reactions in their clients by the fourth interview, but these differences tend to disappear by the eighth interview. On the other hand, differences in the degree of positive subjective reactions elicited from clients are not statistically significant until the eighth interview. From the clients' point of view, the defensive aspects of the relationship appear to develop earlier in therapy than do the positive aspects. Therapists differ in their negative reactions to their

TABLE 7

F RATIOS AND PROBABILITY STATEMENTS FOR AMONG THERAPISTS EFFECTS ON 21 VARIABLES FOR WHICH ANALYSES OF VARIANCE WERE COMPLETED

Variable	<i>F</i> Ratio	<i>p</i>
Client Verbal Behavior Variables:		
Dependence	.98	> .20
Guardedness	7.68	< .01 > .001
Openness	1.08	> .20
Covert Resistance	.86	> .20
Overt Resistance	.71	> .20
Defensiveness	.97	> .20
Relationship Variables:		
Defensive CPRQ—4th interview	4.22	< .05 > .01
Defensive CPRQ—8th interview	2.67	< .10 > .05
Positive CPRQ—4th interview	2.08	< .20 > .10
Positive CPRQ—8th interview	6.88	< .01 > .001
Negative TPRQ—4th interview	3.71	< .05 > .01
Negative TPRQ—8th interview	1.86	< .20 > .10
Positive TPRQ—4th interview	1.55	> .20
Positive TPRQ—8th interview	2.58	< .10 > .05
Change Variables:		
Maladjustment Index	1.17	> .20
Dependency	1.31	> .20
Defensiveness	.92	> .20
Positive Attitudes Toward Self	.26	> .20
Positive Attitudes Toward Others	1.67	> .20
Taylor Anxiety	1.57	> .20
Therapist Posttherapy Rating Scale	.39	> .20

clients by the fourth interview but these differences tend to disappear as therapy progresses. Although not statistically significant, the data suggest that differences in therapists' positive reactions tend to develop as therapy progresses. From the therapists' point of view, the negative aspects of the relationship appear to develop more quickly than the positive aspects. Therapists also appear to differ among themselves in the amount of guarded verbal behavior they elicit from their clients during early interviews. There were no significant differences among these therapists in the changes produced in their clients on the client change variables investigated.

For ascertaining the relationships between therapists' personal characteristics and client reactions, Pearson product-moment correlations were computed between all the therapist variables and five of the client variables. The client variables were the fourth interview Posi-

tive and Defensive CPRQ, Guarded and Defensive verbal behavior in therapy, and the Maladjustment change variable. The values used for computing these correlations were the scores obtained by each therapist on each of the therapist characteristic variables and the mean scores of his four clients on the client variables. The correlations were computed using the total sample of 10 therapists. In those instances in which the correlation coefficients obtained were .30 or above, correlations were computed also for the sample of 6 therapists. None of the correlations met the criterion for statistical significance. This is not surprising since the extremely small samples provide very limited degrees of freedom and thus require a correlation of .63 to be significant at the .05 level. However, six of the eight correlations computed using Edwards' Nurturance Scale yielded correlations of .40 or above, suggesting that the degree to which it is possessed

as a quality by the therapist may have some real relationship to his clients' reactions in therapy. Similarly, the Test of Clinical Judgment, intended to reflect acceptance of the values of others, consistently correlated in the expected direction, although in magnitude the correlations were not statistically significant.

The experimenters feel that the main problem behind the failure to find significant relationships between therapist characteristics and client reactions lies in the approach taken. It is felt that the measures of therapist extratherapy behavior which were obtained were too far removed from their behavior in therapy. The writers now believe that a more profitable approach could be made by investigating therapists' behavior in the interview itself. At this stage of our knowledge it seems rather futile to continue with the investigation of "permanent"

traits of therapists in the hope that they will relate to client behavior.

While none of the therapist personal characteristics studied related to differential reactions of clients, certain patterns of therapist interview behavior seemed to relate qualitatively to their clients' measured behavior. In the course of the investigation the experimenters listened to scores of interviews. From these interviews, the writers agreed that therapists could be grouped along the dimensions of *perceptiveness of client dynamics*, *threat to clients*, and *warmth and friendliness*. It should be made clear, however, that these groupings were relative only to the 10 therapists in this experiment. Characteristics attributed to each group are meant only to depict the principal ways in which members of the group differed from other therapists in the study. Table 8 presents data on se-

TABLE 8
MEAN SCORES FOR THE FOUR CLIENTS OF EACH THERAPIST ON SELECTED VARIABLES
GROUPED ACCORDING TO THE TYPE OF THERAPIST

	Conversational Therapists			Threatening Therapists				Friendly Dynamic Therapists		
	A	B	C	D	E	F	G	H	I	J
Guardedness	1.1	1.8	1.9	4.9	6.7	4.1	7.9	9.3	11.7	2.5
Openness	1.6	2.3	4.8	1.3	3.4	3.7	7.9	7.7	6.0	4.9
Maladjustment	15.8	10.3	14.5	.3	5.5	15.8	4.0	-5.3	2.5	-6.0
Anxiety	21.5	22.5	25.3	25.0	24.8	25.5	24.3	21.5	22.8	26.0
Dependence ^a	6.3	1.5	2.0	-5.5	4.0	3.0	-1.5	2.0	2.0	-1.5
Defensiveness ^a	2.0	-.3	-5.0	-2.0	-1.0	-4.0	0	-1.3	-2.8	0
Attitudes toward others ^a	1.0	1.0	2.0	0	2.0	2.5	1.0	-.5	-1.3	.3
Attitudes toward self ^a	3.0	.3	1.5	-.5	-1.0	1.8	.5	-.5	-.5	1.8
Positive CPRQ ₄ ^b	392.0	303.0	413.0	275.0	390.0	376.0	386.0	317.0	279.0	374.0
Positive CPRQ ₈ ^c	391.0	343.0	419.0	266.0	393.0	398.0	384.0	356.0	279.0	346.0
Defensive CPRQ ₄ ^b	186.0	246.0	172.0	229.0	202.0	209.0	203.0	224.0	170.0	206.0
Defensive CPRQ ₈ ^c	195.0	208.0	185.0	243.0	211.0	178.0	204.0	208.0	170.0	212.0
Positive TPRQ ₄ ^b	107.0	71.0	93.0	101.0	90.0	87.0	108.0	122.0	95.0	100.0
Positive TPRQ ₈ ^c	114.0	60.0	91.0	98.0	96.0	92.0	96.0	125.0	109.0	107.0
Negative TPRQ ₄ ^b	79.0	35.0	62.0	58.0	70.0	46.0	78.0	81.0	92.0	75.0
Negative TPRQ ₈ ^c	74.0	36.0	68.0	65.0	63.0	48.0	77.0	76.0	83.0	74.0

^a Scores represent the difference between pre- and post-therapy scores. Plus scores represent improvement.

^b Fourth interview.

^c Eighth interview.

lected variables for the three groups of therapists discussed below.

Three therapists were characterized as friendly, nonthreatening, and nondynamic, and were labeled "conversational" therapists. They appeared to take the role of "friend to the client." These therapists tried to be as nonthreatening as possible. They tried to communicate to the client that they were on his side. Their therapy seemed to consist mainly of restating content in Reflective Therapy or being reassuring and supportive in Leading Therapy. There was no consistent plan for encouraging clients to discuss problem areas. In fact, "conversational" therapists seemed actively to avoid problem areas at times. One need apparently controlling their behavior was a strong need for approval and acceptance. Clients of these therapists were less guarded in their interview verbal behavior. Consistently, they showed a decrease on the Maladjustment Index of the MMPI from pre- to post-therapy. Strong positive reactions toward the therapy and therapist were reported on the CPRQ, as well as generally low defensiveness reactions.

Four therapists were characterized as non-friendly, threatening, and dynamically oriented. These were labeled "threatening" therapists. They tended to correct the unrealistic perceptions and plans of the client. Through the choice of words, tone, or inflection of voice, client thoughts and actions were directly or indirectly evaluated. These therapists seemed to imply that they "knew the correct answers" to the client's problems or soon would—that they were authorities. They seemed less concerned, or less aware, than the other therapists with the amount of threat involved in their statements. Their interpretations were sometimes quite extreme and without sound basis. They seemed to direct the discussion from one area to another impulsively. "Threatening" therapists tended to be more aggressive and challenging than other therapists. One need apparently controlling their behavior was a need to dominate and control the situation. Three of the four therapists who did not maintain adequate differentiation of the therapies were in this group. Scores for clients of "threatening" therapists were inconsistent from therapist to therapist. However, it is interesting to note that the therapist whose clients were the most defensive, and least positive, on the relationship measures, and least open in the interview, was one in the "threatening" group.

The remaining three therapists were characterized as warm, skillful, and dynamically oriented. These were labeled "warm, dynamic" therapists. While very warm, they tended to be always mindful of the business at hand. In the

reflective therapy, they clarified and restated therapeutically relevant feelings or content without distorting the client's emphasis. In the leading therapy they consistently offered leads more pertinent to the client's dynamics and closer to the client's capacity for dealing with the problem. Although all 10 therapists were generally adequate, the "warm dynamic" therapists tended to approximate more closely the "ideal" therapist as he is frequently described in the literature. The personal needs of these therapists were less obvious and seemed to interfere less with the progress of therapy than the needs of other therapists. They tended to report both more positive and more negative personal reactions than the other therapists. The experimenters fully expected that clients of "warm, dynamic" therapists would earn the most desirable scores on the various measures. However, in virtually every instance their clients failed to do so. Moreover, on some variables they consistently earned the least desirable scores.

It is difficult to account for the apparent poor showing of the clients of the "warm, dynamic" therapists. Why did their clients apparently fail to respond in a manner which would reflect the warmth, acceptance, and skill extended to them? Similarly, it is difficult to account for the apparent superiority of the scores of clients of the "conversational" therapists. The only reasonable explanation seems to be that the variables concerned with client reactions to therapy are actually of a different nature than was originally conceived. It was expected, for example, that the absence of guarded behavior in the interviews would be indicative of a better relationship and greater progress. Noting that the therapists who appeared most skilled seemed to have clients who indulged in relatively more guarded behavior than other clients, a different concept of guardedness emerged. It was found that while clients of the "warm, dynamic" therapists were quite guarded, they were also quite open. In fact, the two kinds of behavior seemed to go together. The confession of an inadequacy, or a socially unacceptable feeling, was usually preceded or followed by guarded statements. It seemed that clients regularly experienced some anxiety in relation to such openness about themselves and needed to cling to some defense in order to allay their anxiety. Conversely, clients of "conversational" therapists showed little tendency to discuss their problems openly. Thus, they had little to be guarded about. The relationship between guardedness and openness seemed to reflect what was occurring in therapy far more meaningfully than either variable considered alone.

The pattern of client scores in relation to therapist interview behavior also

seemed to help clarify the concepts of positive and defensive reactions on the relationship measures. Originally, it was thought that relatively high defensive reactions on the CPRQ would be present only in a threatening therapy. Defensiveness was thought to be detrimental to therapy. It is now felt that a certain amount of defensiveness is probably a necessary concomitant of problem solving arising from the internal resistance of the client to a more realistic evaluation of himself. Relatively high positive scores on the CPRQ were originally conceived of as instrumental to therapy. It is now believed that high positive scores may represent in part dependence on the therapist as an authority figure. This view arose from the fact that therapists assuming benevolent or threatening authority roles achieved the highest scores on the CPRQ at both the fourth and eighth interviews.

The change in maladjustment favored clients of the "conversational" therapists. However, the meaning of this change is not clear. From these data, it might appear that a highly positive relationship is the crucial condition for clients to improve in therapy. However, the writers suspect that the nonthreatening atmosphere created by the conversational therapists resulted in a temporary lessening of anxiety which produced changes in maladjustment scores. It seems more likely that a positive relationship is a necessary but not sufficient condition for effective therapy. As a result of consistent attention to dynamics in a friendly atmosphere, it seems reasonable to expect clients of friendly, dynamic therapists to become more aware of their problems and disturbing feelings. The increase in maladjustment scores at the time the experiment was terminated for such clients may reflect this increased awareness. If effective therapy continued,

this trend might be expected to reverse itself and a decrease in maladjustment scores appear by the end of therapy.

The fact that the friendly, dynamic therapists reported more positive and negative reactions may reflect a higher degree of sensitivity to, and involvement with, their clients. These therapists were apparently more aware of their own reactions and consequently were able to control their own behavior more effectively during therapy.

On the basis of the scores obtained by the three types of therapists described, it is suggested that a more significant index of the effectiveness of therapy may be the relationship between guarded and open behavior. Similarly the relationship between clients' positive and defensive subjective reactions may be an index to a truly therapeutic relationship. If there is too much guardedness in relation to openness and/or too much defensiveness in relation to positiveness, therapy is probably not proceeding optimally. Low scores on guardedness with little openness may indicate that little problem solving is occurring. High scores on guardedness with little openness may indicate that the therapist is unduly threatening. High scores on positiveness with little defensiveness suggest that the relationship may be satisfying to the client but not necessarily therapeutic. It should be pointed out that these proposed relationships apply only to the early stages of therapy. In later stages the optimal relationships of these scores may be quite different.

Differences Resulting from Interaction of Therapist with Therapy

Three of the 21 analyses revealed statistically significant interaction F ratios as indicated in Table 9. These differences appear in the analyses of the client relationship measures. The personal qualities which therapists invested

TABLE 9

F RATIOS AND PROBABILITY STATEMENTS FOR INTERACTION (THERAPIST \times METHOD) EFFECTS ON
21 VARIABLES FOR WHICH ANALYSES OF VARIANCE WERE COMPLETED

Variable	F Ratio	p
Client Verbal Behavior Variables:		
Dependence	1.60	> .20
Guardedness	2.12	< .20 > .10
Openness	.38	> .20
Covert Resistance	1.09	> .20
Overt Resistance	.64	> .20
Defensiveness	1.15	> .20
Relationship Variables:		
Defensive CPRQ—4th interview	7.92	< .01 > .001
Defensive CPRQ—8th interview	6.61	< .01 > .001
Positive CPRQ—4th interview	2.98	< .10 > .05
Positive CPRQ—8th interview	9.56	< .001
Negative TPRQ—4th interview	.99	> .20
Negative TPRQ—8th interview	.40	> .20
Positive TPRQ—4th interview	.73	> .20
Positive TPRQ—8th interview	.21	> .20
Change Variables: ^a		
Maladjustment Index		
Dependency	1.22	> .20
Defensiveness	.79	> .20
Positive Attitudes Toward Self	.16	> .20
Positive Attitudes Toward Others	.87	> .20
Taylor Anxiety Scale	.48	> .20
Therapist Posttherapy Rating Scale	.60	> .20
	.42	> .20

^a All analyses were made on the difference between pre- and post-therapy scores except for the Taylor and Posttherapy Rating Scale where posttherapy scores were used.

in each of the therapies affected the degree of defensive reactions of clients. Such effects are pronounced by the fourth interview and persist at least through the eighth interview. Client positive reactions tend to vary in the same way. By the eighth interview, differences in client positive reactions are clear-cut. However, therapist reactions do not show similar interaction effects. None of the client verbal behavior variables or variables intended to reflect change during therapy revealed any significant interaction effects.

Therapists interacting with the therapy they are administering apparently create different effects in the positive and defensive aspects of the relationship from clients' points of view. The defensive aspects of the relationship tend to become less related to therapists as persons but remained related to the interaction

of therapist with the type of therapy administered. On the other hand, the positive aspects of the relationship become more clearly related to therapists as individuals as well as to therapists interacting with the type of therapy administered as therapy progresses.

Inspection of the raw data suggests an explanation of these changes. On the defensive CPRQ, total means for each therapist tend to converge toward the total mean for all therapists between the fourth and eighth interviews. At the same time, the difference between the reflective and leading treatment means for each therapist tends to become greater between the fourth and eighth interviews. This suggests that as therapy progresses clients tend to react more selectively to their therapists' behavior. Changes on the positive CPRQ result primarily from changes in the error term

in the F ratios. At the fourth interview, the error term accounts for about one-third of the total variability, while at the eighth interview it accounts for only about one-eighth of the total variability. In other words, the scores of pairs of clients for each therapist in each treatment condition became more alike between the fourth and eighth interviews, while differences among averages of the scores for each therapist and each therapy tended to remain the same. This suggests that clients' subjective reactions tend to become less influenced by their habitual interpersonal sets and more by factors within the therapeutic situation.

A generalized fear of intimate interpersonal situations may be one of the more important of these sets. As this generalized fear is diminished, the client becomes more discriminating in his reactions. Because the client is reacting more discriminatingly to the stimuli within the situation, a more favorable set of circumstances for therapeutic change would appear to be developing.

Several factors may be related to these interaction effects. Therapists' training and experience may influence the effectiveness with which they administer different therapies. In this regard, it is interesting to note that the two therapists with the most extensive training and experience with leading forms of psychotherapy obtained the highest defensive CPRQ scores in the leading treatment at both the fourth and eighth interviews. Therapists' own dynamics and personal value systems may relate to the way in which they make use of a type of therapy. The types of roles assumed by therapists may vary with the type of therapy offered. This study provides no strong clues toward determining which factors are significantly related to these differences.

General Discussion

This study indicates that clients' views of the therapeutic relationship depend on the interaction of the clients' own dy-

namics, the kind of therapy administered, and individual characteristics of therapists. This idea has been discussed frequently in the literature. On the other hand, the idea that one type of therapy is good for all clients and that any therapist can be effective with a given type of therapy has also been suggested. The present data tend to support the former rather than the latter view.

The implications of these results seem of considerable importance. Extensive research is needed to define differential client reactions to different therapies and to different therapist characteristics. Research is also needed to define the client pretherapy characteristics related to differential client reactions, as well as research relating client reactions during therapy to therapeutic outcomes.

The consistently significant findings on client guarded or defensive verbal behavior and the client positive and defensive relationships measures also suggest that these measures are sensitive to some of the effects of therapy and therefore merit further study. This observation may have broader implications, however, for research in this area. If one wants to measure what happens in therapy, behavior in therapy or reactions to the therapy situation specifically may be one of the most sensitive areas of measurement.

This study clearly illustrates the value of multivariate experimental designs with measures taken at different times when studying as complex an area as psychotherapy. The writers firmly believe that major advances in the effective use of psychotherapy will come most quickly from carefully designed and executed research. Only in this way will therapists become able to administer therapy based on a body of verified

knowledge rather than finding it necessary to rely primarily on individual clinical experience.

CONCLUSIONS

The view that a leading and a reflective type of therapy produce different effects on clients was slightly supported. During the first eight interviews, these therapies did not produce significantly different effects on the defensive or positive aspects of the relationship from the clients' points of view nor in the positive or negative aspects of the relationships from the therapists' points of view. The therapies did not differ significantly in the extent of defensive, dependent, open, covertly resistive, or overtly resistive verbal behavior elicited from clients during the first four interviews. The therapies did not differ significantly in the amount of client change during therapy in dependence, defensiveness, maladjustment, positive attitudes toward self, positive attitudes toward others, or anxiety. Of 21 variables explored, client guarded verbal behavior and the Therapists' Posttherapy Rating Scale were the only ones showing a statistically significant difference.

The view that pretherapy characteristics of clients relate differentially to client reactions to therapy in reflective and leading types of therapy is partially supported. There were differences between the two types of therapy in the manner in which pretherapy defensiveness, and pretherapy aggressive need related to verbal defensive behavior of clients in therapy. There were also differences between the two types of therapy in the manner in which pretherapy needs to be deferent and to be autonomous related to client subjective defensive reactions to therapy. None of the differences in the other 20 sets of corre-

lations were statistically significant. The differences between treatments reflect the following findings. Clients who were more defensive when they entered therapy tended to behave more defensively when they were in leading therapy. There was no such relationship under reflective therapy. There were tendencies for clients who entered therapy with more aggressive need to behave more defensively in the leading treatment, and less defensively in the reflective therapy. Clients who entered therapy with more need to be deferent to others felt more defensive in the leading treatment. No such relationship was apparent in the reflective treatment. Clients who entered therapy with a strong need for autonomy tended to feel less defensive in leading therapy. Autonomy need seemed unrelated to defensive feelings in the reflective treatment.

The view that individual therapists create different effects on their clients independent of the type of therapy given is partially supported. Therapists in this study differed significantly in the defensive and positive feelings they elicited from their clients during the first eight interviews. They also differed significantly in the extent of guarded verbal behavior exhibited by their clients in the first four interviews. However, they did not differ significantly in the extent of defensive, dependent, open, covertly resistive, or overtly resistive verbal behavior elicited from clients in the first four interviews. They did not differ significantly in their view of the positive or negative aspects of the relationship during the first eight interviews. They did not differ significantly in the extent of change produced in their clients in maladjustment, dependence, defensiveness, positive attitudes toward self, positive attitudes toward others, or

anxiety during the course of therapy. They did not differ significantly in their evaluation of the extent of change produced in their clients as a result of therapy.

The view that selected therapist characteristics are related to the kinds of relationships established, the amount of defensive or guarded verbal behavior elicited from clients, and the amount of change in adjustment produced in clients is not supported. Therapists' ability to enter the phenomenological field of another, sympathetic interest, acceptance of the value system of others, social stimulus-value to associates, need to aggrandize the self, and aggressiveness did not correlate significantly with any of the five dependent variables examined.

The view that the interaction of the therapist as an individual and the type of therapy he is employing affects clients is partially supported. The way in which therapists used or molded a type of ther-

apy had effects upon clients. Clients felt significantly more defensive or more positive in one type of therapy with individual therapists than did other clients for the same therapist in the second type of therapy. For some therapists the increased defensive or positive feelings were in the leading treatment while for other therapists the reactions elicited were greater in the reflective treatment. The clients in this study did not differ significantly as a result of the interaction effect (therapist \times method) in the extent of guarded, defensive, dependent, open, covertly resistive, or overtly resistive verbal behavior which they manifested. They also did not differ significantly as a result of the interaction effect in the extent of change they exhibited in maladjustment, dependence, defensiveness, positive attitudes toward self, positive attitudes toward others, anxiety, or therapists' evaluation of change as a result of therapy.

REFERENCES

1. ASHBY, J. D. The effect of a reflective and of a leading psychotherapy on certain client characteristics. Unpublished doctoral dissertation, Pennsylvania State Univer., 1956.
2. BORDIN, E. S. The implications of client expectations for the counseling process. *J. clin. Psychol.*, 1946, 2, 17-21.
3. BOWN, O. H. An investigation of therapeutic relationship in client-centered psychotherapy. Unpublished doctoral dissertation, Univer. of Chicago, 1954.
4. CURRAN, C. A. *Personality factors in counseling*. New York: Grune and Stratton, 1945.
5. DAULTON, M. J. A study of factors relating to resistance in the interview. Unpublished master's thesis, Ohio State Univer., 1947. Cited in F. P. Robinson, *Principles and procedures in student counseling*. New York: Harpers, 1950.
6. DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
7. EDWARDS, A. *Manual, personal preference schedule*. New York: Psychological Corp., 1954.
8. FIEDLER, F. E. Factor analyses of psychoanalytic, nondirective, and Adlerian therapeutic relationships. *J. consult. Psychol.*, 1951, 15, 32-38.
9. FORD, D. H. An experimental comparison of the relationship between client and therapist in a reflective and a leading type of psychotherapy. Unpublished doctoral dissertation, Pennsylvania State Univer., 1956.
10. FRANK, J. D. Experimental studies of personal pressure and resistance. *J. gen. Psychol.*, 1944, 30, 23-41.
11. FROMM-REICHMANN, FRIEDA. *Principles of intensive psychotherapy*. Chicago: Univer. of Chicago Press, 1950.
12. GALLAGHER, J. J. MMPI changes concomitant with client-centered therapy. *J. consult. Psychol.*, 1953, 17, 334-338.
13. GALLAGHER, J. J. The problem of escaping clients in nondirective counseling. In W. U. Snyder (Ed.), *Group report of a program of research in psychotherapy*. State College, Pa.: Pennsylvania State Univer., 1953.
14. GIBSON, R. A factor analysis of measuring

- changes following client-centered psychotherapy. Unpublished doctoral dissertation, Pennsylvania State Univer., 1953.
15. GILLESPIE, J. F. Verbal signs of resistance in client-centered therapy. In W. U. Snyder (Ed.), *Group report of a program of research in psychotherapy*. State College, Pa.: Pennsylvania State Univer., 1953.
 16. GUERNEY, B. G., JR. Client dependency, guardedness, openness, and resistance in a reflective and in a leading psychotherapy. Unpublished doctoral dissertation, Pennsylvania State Univer., 1956.
 17. GUERNEY, L. F. Differential effects of certain therapist characteristics on client reactions to psychotherapy. Unpublished doctoral dissertation, Pennsylvania State Univer., 1956.
 18. HAIGH, G. Defensive behavior in client-centered therapy. *J. consult. Psychol.*, 1949, 13, 181-189.
 19. HOGAN, R. A. A measure of client defensiveness. In W. Wolff and J. A. Precker, *Success in psychotherapy*. New York: Grune and Stratton, 1952.
 20. MOONEY, R. L. Surveying high school students' problems by means of a problem check list. *Educ. Res. Bull.*, 1942, 21, 57-69.
 21. MOWRER, O. H. *Psychotherapy, theory, and research*. New York: Ronald Press, 1953.
 22. NAVRAN, L. A rationally derived MMPI scale to measure dependence. *J. consult. Psychol.*, 1954, 18, 192.
 23. ROGERS, C. R., *Counseling and psychotherapy*. Boston: Houghton Mifflin, 1942.
 24. ROGERS, C. R. *Client-centered therapy*. Boston: Houghton Mifflin, 1951.
 25. ROGERS, C. R., & DYMOND, ROSALIND. *Psychotherapy and personality change*. Chicago: Univer. of Chicago Press, 1954.
 26. SEEMAN, J. A study of the process of non-directive therapy. *J. consult. Psychol.*, 1949, 13, 157-168.
 27. SIEGEL, S. Certain determinants and correlations of authoritarianism. *Genet. Psychol. Monogr.*, 1954, 49, 187-229.
 28. SNEDECOR, G. W. *Statistical methods*. Ames, Iowa: Iowa State College Press, 1946.
 29. SNYDER, W. U. An investigation of the nature of nondirective psychotherapy. *J. genet. Psychol.*, 1945, 33, 193-223.
 30. SNYDER, W. U. *Group report of a program of research in psychotherapy*. State College, Pa.: Pennsylvania State Univer., 1953.
 31. STRUPP, H. An objective comparison of Rogerian and psychoanalytic techniques. *J. consult. Psychol.*, 1953, 19, 1-7.
 32. TAYLOR, J. A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
 33. TUCKER, J. E. Measuring client progress in client-centered psychotherapy. In W. U. Snyder (Ed.), *Group report of a program of research in psychotherapy*. State College, Pa.: Pennsylvania State Univer., 1953.

(Accepted for publication April 5, 1957)

APPENDIX

THE CLIENT'S PERSONAL REACTION QUESTIONNAIRE

DIRECTIONS

During the process of personal counseling, people have many different feelings and reactions. We know that these reactions are sometimes negative, sometimes positive, and often mixed. Your responses to the following questionnaire will help us understand people's reactions to personal counseling. This will have nothing to do with your counseling. It will be *completely confidential*. Neither your counselor nor his superiors will be informed of your responses. Approximately 15 minutes are required to complete it.

There are five possible responses to each of the items in the questionnaire.

- 1 not characteristic
- 2 slightly characteristic
- 3 moderately characteristic
- 4 quite characteristic
- 5 highly characteristic

Put a circle around the responses most representative of your present feelings. Your feelings

may have been different in the past and may be different in the future. We are interested in your feelings right now at this point in your counseling experience. Be sure to put a circle around *one* response for *each* item. Do not spend too much time on any one item.

POSITIVE ITEMS¹

1. I'm pleased with my counselor's interest and attention.
2. I wish I felt as sure of myself in all social situations as I do here.
3. I like our sessions even when I can't think of anything to say.
4. I remember and "chew over" things that my counselor says.
5. I have a very warm feeling toward my counselor.
6. I wish I had some friends who were as understanding as my counselor.
7. I am usually eager to hear what my counselor has to say.
8. I often feel in a better mood after an interview.
9. I sometimes feel like letting my counselor know what a nice person I think he (she) is.

¹ Grouped for the reader's convenience.

10. I wish I could feel other people respected and liked me as much as my counselor does.
11. I feel comfortable talking with my counselor.
12. This is one of the few situations I've ever been in that I didn't worry very much about what the other person thought of me.
13. I wish I were more like my counselor.
14. I feel that my counselor regards me as a likable person.
15. Many of the things my counselor says just seem to hit the nail on the head.
16. I experience a certain relief after telling my counselor something.
17. My counselor's attitude gives me hope I can get something out of this.
18. I feel that the counselor really likes to spend the counseling session with me.
19. I feel my counselor is really anxious to help me solve my problems.
20. It's easier for me to talk with this counselor than with most other people.
21. I think I could criticize or get angry at my counselor and he (she) wouldn't resent it.
22. My counselor must be one of the best ones here.
23. My counselor's understanding of me is encouraging.
24. I usually feel the interviews have been worth while.
25. I really get "wrapped up" in what's going on in the counseling session.
26. I wish I had asked for this kind of help sooner.
27. I can talk about most anything in my interviews without feeling embarrassed or ashamed.
28. I wish I could spend more time with the counselor.
29. I am gaining more respect for psychology as a result of my experiences in counseling.
30. The things my counselor says and does give me confidence in him.
31. I seldom feel the counselor has misinterpreted what I have said or done.
32. I know the counselor understands me even when I don't express myself well.
33. I have the feeling here is one person I can really trust.
34. I would like to behave toward other people more like my therapist behaves toward me.
35. I look forward to talking with my counselor.
36. I feel sure that my counselor would take anything I could say or do without getting upset.
37. I'm pleased with the progress I've made since beginning these interviews.
38. I'm glad this particular counselor was assigned to me.
39. The counselor is a warm and friendly person.
40. I think my counselor really sympathizes with my difficulties.
7. It takes a long time in an interview for me to get started talking about important things.
8. It seems to me there should be an easier and quicker way than this to solve my problems.
9. Many of the things we talk about don't seem to be related to my problems.
10. I sometimes feel like leaving before the interview hour is over.
11. Sometimes after the counselor says something I just can't think of anything else to say for awhile.
12. I sometimes hesitate to tell my counselor what I am really thinking.
13. If I had someone else as a counselor, I would probably feel freer to discuss my problems.
14. These interviews seem like a waste of time to me.
15. Sometimes I feel like I'm being criticized during the interviews.
16. It doesn't make much difference to me whether my counselor likes me or not.
17. I frequently find it difficult to think of things to say.
18. I get irritated at some of my counselor's comments.
19. I spend very little time thinking about these interviews when I'm not here.
20. It was kind of unnecessary for me to start this counseling because my problems really aren't very major.
21. I just don't know what to do or say in the interviews that would help.
22. When I'm in the counseling session I sometimes forget things I had meant to tell my counselor.
23. If my counselor understood me better I could make more progress.
24. I try to justify my actions so the counselor will see why I behaved the way I did.
25. There are some things which I don't yet feel ready to go into with my counselor.
26. I feel a need to keep the conversation moving during the counseling hour.
27. I can't see where my counselor has done much to help me solve my problems.
28. I don't know exactly why, but I feel nervous about coming to the counseling hour.
29. Sometimes it's hard for me to pay attention to what the counselor is saying.
30. I know what my problems are, but I don't know what to do about them.
31. I carefully organize what I'm going to say in the next counseling session.
32. The counselor's looking at me all the time makes me uncomfortable.
33. I sometimes wish we were talking about something different than what we're talking about at the moment.
34. If I would take a couple of hours a week to think about these things on my own, I could probably accomplish as much as I do in these interviews.
35. I feel like my counselor wants me to tell him a lot more than I am telling him.
36. I'm afraid to express my real feelings in these sessions.
37. I don't think I have as many problems as other people in counseling.
38. I sometimes resent the counselor's attitude toward me.
39. I sometimes feel like I'm being put on the spot.
40. I feel that my counselor has me classified or categorized as some kind of "case."

DEFENSIVE ITEMS

1. It would be helpful if I could write things down and bring them in to the counselor to discuss.
2. I doubt if many people get much help out of these interviews.
3. Sometimes the counselor seems to twist around the things I say to mean something different than what I intended.
4. It's hard for me to talk about myself.
5. I sometimes feel like calling this whole thing off.
6. I've told the counselor a lot but he (she) still hasn't given me much help.

THE THERAPIST'S PERSONAL REACTION QUESTIONNAIRE

DIRECTIONS

During the process of personal counseling, counselors have many different feelings and reactions. These reactions are sometimes negative,

sometimes positive, and sometimes mixed. Leaders in various schools of therapy seem to agree that having varied feelings and reactions toward clients is not undesirable as long as the coun-

selor recognizes and understands them. In fact, they may provide additional sensitive evidence about the meaning of a client's communications. We are interested in learning what these feelings are and how they change.

Your responses will have nothing to do with your counseling. They will not be made available in identifiable form to your supervisor or anyone else except the researchers.

There are five possible responses to each of the items in the questionnaire.

- 1 not characteristic
- 2 slightly characteristic
- 3 moderately characteristic
- 4 quite characteristic
- 5 highly characteristic

Put a circle around the responses most representative of your present feelings. Your feelings may have been different in the past and may be different in the future. We are interested in your feelings right now at this point in your counseling experience with this client. Be sure to put a circle around *one* response for *each* item. Do not spend too much time on any one item. The numbers may appear in different orders before each item, but they will always signify the same response. Some of the items are written comparing this client with other clients you have had. Try to respond to the items in terms of that comparison, if possible.

When you have completed the questionnaire, return it to the envelope and drop the sealed envelope in the box marked TESTS in the waiting room. The test *must* be completed before your next interview with this client.

POSITIVE ITEMS²

1. I like this client more than most.
2. I have a more warm, friendly emotional reaction toward this client than others.
3. I feel sure that this client would not want to change therapists if given the chance.
4. I am seldom in doubt about what the client is trying to say.
5. I think about this client more often between meetings than others.
6. This client seems to appreciate my efforts.
7. In general, I could not ask for a better client.
8. I'll miss having these interviews a little when the client decides to terminate.
9. I prefer working with this client more than others I've worked with.
10. If I rated all of the clients I've worked with so far in my career in terms of the satisfaction I've gotten out of them, this client would receive a high rating.
11. I think this client is trying harder to solve his (her) problems than others I've had.
12. I sometimes wish the therapy sessions with this client would not end so soon.
13. I am more confident this client will work out his problems than I've been with others.
14. I can usually find significant things to respond to in what the client says.
15. I think I'd like this client socially if I had met him first in that capacity.

² Grouped for the reader's convenience.

16. We get into more important material than is frequently the case with other clients.
17. It's easier for me to see exactly how this client would feel in the situations he describes than it is with others.
18. I sometimes feel like congratulating this client for something he has done.
19. I'm usually more absorbed in what this client is doing or saying than with others I've had.
20. Responses to what this client is saying come more "easily" than with others.
21. We may have our ups and downs, but underneath it all I think the client has confidence in me.
22. Therapy with this client is a more rewarding experience for me than with many others I've had.
23. When things are not going well for the client I feel upset too.
24. I usually have a good feeling about interviews with this client.
25. I think I'm doing a pretty competent job with this client.
26. I think we have a pretty relaxed, understanding kind of relationship.
27. I feel more comfortable in the therapy sessions with this client than with others I've had.
28. I find it easier to understand and communicate with this client than with others.
29. As compared to others, I'm pretty "wrapped up" in this client and in trying to help him.
30. If I had to leave here for some reason before this client was finished, I'd try very hard to see that he was assigned to "just the right therapist."
31. I'm glad this particular client was assigned to me.
32. I get anxious about what to do or say with this client less frequently than with others I've had.
33. I look forward to my interviews with this client more than with others I've had.
34. Similarities between my own emotional experiences and some of this client's make me feel a little closer to him (her) than to others I've had.
35. Relationships like this are bright spots in my schedule.

NEGATIVE ITEMS

1. I'm usually relieved when the interviews are over.
2. I can't get this client to open up.
3. I get pretty bored in some of these interviews.
4. Sometimes I get pretty tense during the interviews.
5. I seldom feel that we have accomplished something in the interviews.
6. I would like to be able to feel more warmth toward this client than I now feel.
7. I don't feel the client is making as much use of therapy as he could.
8. I don't particularly enjoy my hour with this client.
9. I can't get close to this client.
10. I sometimes wonder if another therapist might not get further with this client than I.
11. I disagree with this client about some basic matters like religion, morality, etc.
12. It is really an effort to "stay with" this client.
13. This client "hits me where it hurts" sometimes.
14. I find it harder to remember what has been covered in previous interviews with this client than with others.
15. In comparison with other clients, I find it hard to get involved with this client's problems.
16. I feel in need of help with this case.
17. I can't seem to get very interested in this client.
18. I seem to tire more quickly when I work with this client than with others.
19. I sometimes feel "pushed" by this client.
20. Sometimes I feel pretty frustrated in our interviews.
21. I have to exert more self-control and self-restraint with this client than with most.
22. I doubt if any counselor could do much for this client.

23. I get pretty discouraged at times about this one.
24. I feel pretty ineffective with this client.
25. I prefer working with this client less than others I've worked with.
26. I can't help but be annoyed to some extent by some of this client's behavior.
27. I'm "flying blind" with this client.
28. Sometimes I resent the client's attitude.
29. It's hard to know how to respond to this client in a helpful way.
30. In comparison with other clients, it's hard for me to put myself in this client's place.
31. I am sometimes at a loss as to how to respond to this client.
32. I don't think this client will stand out in my pleasant memories of cases.
33. Sometimes I have to show more sympathy and acceptance than I really feel toward this client.
34. Sometimes I wish some other therapist had this client.
35. The hour often seems to be dragging on with this client.

CLIENT VERBAL BEHAVIOR CATEGORIES

GENERAL RULES

1. A client response is a client statement or 15 seconds of silence between two therapist statements. If a client's statement is longer than 15 seconds it is divided in 15-second units. The last part of a client statement between a previous 15-second portion and the time the therapist speaks is categorized even though it may be less than 15 seconds long, unless it is a pause.

2. The category "N" is used for the client only when there is nothing in the client's response to call for a classification anywhere else.

3. If a statement cannot definitely be placed in a category other than "N," because it is not complete at 15 seconds when the recording machine is stopped, put it in that "N" category. But its meaning is allowed to have bearing on the categorization of the next 15-second portion of the client's response. If, on the other hand, enough of the client's statement is heard to put it definitely in a category, do so even though the sentence may not be complete.

4. The same sentence cannot be categorized in two places (unless it is an Interruption, Blocking, or Change of Topic), but more than one category may be used for the same 15-second interval.

5. If in conflict between two or more categorizations of the same client sentence, place it where the least inference is required.

6. We are interested in the variables being measured here only as they are manifest in the counseling situation itself. Client statements which indicate only that the client is or has been, for example, guarded or dependent outside of the therapy hour do not fall under the Guardedness or Dependency categories here.

DEPENDENCY

Briefly defined, "Dependency" is the extent to which the client asks the therapist for his opinions, or advice, information, evaluation, instruction, or demonstrates a need for structuring from the therapist. Also it includes the degree to which the client places responsibility for progress or outcome of counseling on the counselor rather than accepting it himself.

Statements to be categorized as Dependency are statements in which the client:

1. Asks for help, opinions, advice, solutions, information, judgments, evaluations, instructions from the therapist. Asking the therapist for elaborations on his statement is also included here.

2. Shows by his statement that he is putting the responsibility for providing the above-mentioned things on the therapist, or he puts the responsibility for progress or outcome of therapy on the therapist. This can be shown indirectly through such statements as "You won't be able to figure out what's wrong with me unless I keep talking."

Note: Statements pertaining to appointment time, and asking for cigarettes, match, etc. are not categorized here.

Rhetorical questions (e.g., "Did I tell you about . . .") are not categorized here. Some pause or other indication that the client expects an answer is necessary for client questions of this sort to be categorized as "Dependency."

On the other hand, there may be statements which are clearly questions being asked of the therapist, although they may not be put in question form. For example, "I wonder why that would be (pause)." If such a statement seems aimed at the therapist, it should be categorized as "Dependency."

The following statements are some examples to clarify the category of Dependency:

1. "I didn't expect this to be easy or direct, but it does enter my mind, just what can be accomplished here and how?" . . . D. If a similar statement had cast more doubt on the worth of the therapist or therapy a Covert Resistance categorization would be warranted in addition to a Dependency categorization.

2. Therapist: "She didn't want to." Client: "You mean my sister?" . . . N. A question which calls for clarification of an ambiguous therapist statement does not receive a Dependency categorization. If the therapist's statement is judged clear in meaning and the client nevertheless requests clarification this may receive a Covert Resistance categorization.

3. Therapist: "Maybe you could do some thinking about that." Client: "In what way?" . . . D. The client is asking for instruction.

4. "When do you think a person should start thinking about getting married?" . . . D. This is a request for therapist's opinion.

OPENNESS

"Openness" is defined as: the extent to which the client freely discusses his problems, deviations from the "normal," his culturally frowned-upon traits, behavior, and motivations; and, in general, his willingness to expose himself to potential criticism and change, particularly his willingness to discuss thoroughly those areas which seem most threatening. He does this without at the same time qualifying, hedging, and engaging in defensive verbal maneuvers.

In a way this category is the opposite of the "guardedness" category. When the client's statement does not exhibit any guardedness in a discussion area in which a clinician would often expect a person to be guarded, the client's statement is categorized under "Openness." This category is meant to reflect the ability or willingness of a client to "open himself up" to the counselor; to expose himself to possible criticism, to expose himself to the prospect of modifying his self concept, without at the same time feeling the need to defend himself as he is to his own or to the counselor's eyes. A confidence in the therapist's understanding, noncriticalness and trustworthiness is presumed to underlie such openness; i.e., the statements that are considered "open" are not such as one would tell to one's critics or even new-found friends.

1. Unqualified³ admission to a problem,⁴ deficiency, in-

³ An admission to a problem is "qualified" when it is coupled with expression of uncertainty (e.g., "maybe," "perhaps," "probably," etc.); the term "I think," however, is not considered to indicate uncertainty). Qualification may also exist in any minimization of the problem in terms of the extent or severity of the problem. Such minimization is, of course, a different thing than clarification and specification, and the coder must make a decision as to whether the client is being cau-

adequacy, undesirable characteristic, trait, behavior pattern or act, feelings or attitudes.

2. Unqualified statements pointing to personal deviation from the norm in a culturally or personally undesirable direction.

3. Unqualified statements that admit to possession of culturally or personally undesirable characteristics, traits, behavior, feelings, or attitudes.

Note: Simple acceptance of a therapist's statement placing the client in an undesirable light is not scored "Open." It is only when the client proceeds in such a way as to place his statement under any of the criteria listed above that his statement is categorized "Open."

A simple statement of a problem qualifies as an "Open" response if it is unqualified.

A description of a particular conflict is not necessarily "Open." But an unqualified admission of having important unresolved contradictions or irrationalities within oneself is categorized "Open."

The following statements are some examples to clarify the categorization of "Openness":

1. Client (speaking about husband or father): "As far as really deep feelings—I have none for him." . . . O. This is an unqualified admission to culturally unacceptable attitude or feeling.

2. "My conversational ability is pretty weak. I can't carry on a long continuous conversation." . . . O. This is an unqualified admission of an inadequacy.

3. Therapist: "That seems kind of contradictory." Client: "When I think about it, that's true; as far as things have gone in the past I have no reason to feel inferior." . . . O. The client accepts and elaborates upon therapist statement pointing to important contradiction or irrationality in client.

4. "Most of the time I'm worried about people—what they're thinking of me." . . . O. This is an unqualified admission of a psychological problem.

5. Therapist: "And you feel like he's thinking only of himself." Client: "Not exactly. I don't blame him particularly. I guess I do in a way. I don't want to, but I do. I want him to accept me and my needs." . . . O. This is an admission to a feeling toward which a personal distaste is made clear.

GUARDEDNESS

Guardedness is defined as the extent to which the client exhibits wariness and hedging in regard to presenting and working on his problem, admitting to faults, and exposing himself to potential criticism and change. This includes self-stimulated denial, or minimization, of his problems or his deviations from the "normal"; and denial of culturally undesirable feelings, traits, and motivations. It also includes the need to justify himself or his actions to the therapist and expectations or anticipations of criticism from the therapist.

A key question for the judge to ask himself in listening to a client's statement with reference to this category is whether the client is in any way engaged in "protecting" himself from potential criticism or potential change in his self concept. Some of the cues to listen for are presented below.

1. Statements denying, qualifying, minimizing, or belittling the extent of a problem or the existence of

tious and guarded, or merely giving some definitive information to his counselor. When the case is ambiguous to the coder he should not categorize it as "Openness" (or Guardedness) but as "None." A qualification must occur in the same 15-second interval as a statement in order for it to disqualify the statement as "open." If an admission is accompanied (i.e., in the same 15-second interval) by anything that calls for a Guardedness categorization it is not categorized as Openness. These categories are mutually exclusive in the same client response.

4 The word "problem" is defined in the section on "Guardedness."

one.⁵ The denial is not in response to a question or statement of the therapist or any other particular person.

2. Statements pointing to nondeviation from the "norm," "average," "everyone," "other," etc., either as a person in general or in some particular aspect of thought, feeling, or behavior (e.g., "I guess we all have a tendency to talk in circles."). If such a statement is purely descriptive and factual or is in the nature of a complaint, rather than being something which is *comforting to the client*, it does not receive a Guardedness categorization.

3. Statements denying possession of an undesirable characteristic, trait, feeling, attitude, or denying an undesirable act or motivation. This category is not used when the characteristic has been attributed to the client by the therapist or some specific other person.

4. Statements in which the client attempts to justify an act, thought, feeling, statement, etc. to the therapist. "Justify" here means to hold forth one's behavior as just, right, warranted; to declare oneself guiltless; absolve or acquit oneself; to attempt to show satisfactory excuse or reason for something that is culturally or personally undesirable. The reason that is offered by the client is usually one that is primarily "outside" of the self, i.e., the undesirable thing is a result of the behavior of others or circumstances. This category does not apply if the client has been asked to justify himself by the therapist or some specific other persons.

5. A statement indicating that the client might be anticipating a critical or differing thought or statement from the therapist. (For example, "That's the way I think about it, you may think differently, I don't know"). Such phrases as "sound to you" or "look to you" from the client to the therapist should alert the coder to the possibility that such a statement is an anticipation of difference or criticalness between the client and the counselor.

Some such statements could serve to beat the therapist to the punch. "All this must sound foolish to you." Here the expectation by the client of a critical attitude on the therapist's part is overt. In some cases the anticipation of a critical attitude is only implied by the client's overt acceptance of the blame or fault-finding he expects from the therapist. He will thus make a statement taking fault or responsibility upon himself, while he is in reality rejecting the blame or criticism. Sometimes this takes the form of a mutually qualified acceptance of a statement by the therapist or some other person which places the client in a bad light. Admitting to blame does not, of course, automatically indicate that the statement should be categorized as Guardedness. Such a statement may belong in the N or OP category depending on the way it is said and:

1. The extent of qualification. As a rule, in doubtful cases, do not categorize as G an unqualified statement accepting blame (e.g., "I'm being foolish.") Do not categorize as G a statement containing a single qualification (e.g., "Maybe I'm being foolish.") Do categorize as G such a statement which contains a plural number of qualifications (e.g., "I guess maybe I'm being foolish.")

2. The nature of the elaboration on the statement that is found within the same client response. As a rule, in doubtful cases: (a) the latter part of the client's response is considered more important—if the

5 "Problem" is broadly defined here as anything which is culturally frowned-upon, or which bothers the client personally. To qualify under the latter the client must make his personal distaste clear. Physical symptoms or manifestations of problems (e.g., headaches, crying, sweating, shaking) stated only as a physical problem are not considered to fall under the definition of "problem." But psychological symptoms (e.g., nervousness, feeling depressed, etc.) are considered to be "problems."

acceptance of fault follows the blaming of outside circumstances or other people, it should probably not be categorized as G. On the other hand, if the blaming of outside factors follows self-blame, this should probably be categorized as G; (b) the more detailed and specific part of the client's response is considered more important and the vague, more general, part considered to be less important. Thus, if a client states he is at fault in some general way, but criticizes another person or a circumstance in a more specific way, the statement is likely to be a Guardedness statement.

Note: A simple rejection of a therapist's statement, whatever it may have been, is not categorized as G. It is only when the client proceeds in such a way as to place his statement under any of the criteria listed above that his statement is categorized as Guardedness.

The following statements are some examples to clarify the categorization of Guardedness:

1. "I wonder if I have any problems. Maybe it's just that I think I have problems. Maybe that's all there is to the whole thing." . . . G. This is minimization of the problem.

2. "I guess everybody feels that way about something." . . . G. This is self-stimulated pointing to nondeviation from the norm.

3. "I was proud of the medal, and I showed it to everyone, as every successful athlete would." . . . G. The client is pointing to nondeviation from the norm in regard to the culturally frowned-upon trait of pride.

4. "I don't like to visit my family because when I have too much work to do it bothers me." . . . G. This is justification.

5. Therapist: "You feel inferior to them." Client: "Maybe that's true. I don't know. Anyway, they kept talking about things with which I wasn't familiar, which I thought was very inconsiderate of them." . . . If the manner in which the statement is said is consistent with such a categorization, this doubly qualified acceptance is put in the G category as an overt acceptance of a therapist's statement placing the client in an undesirable light which is really emotionally rejected by the client.

6. Therapist: "You feel inferior to them." Client: "No, it wasn't that. It was just that they kept talking about things with which I wasn't familiar, which I thought was very inconsiderate of them." . . . N. A rejection of a therapist's statement placing the client in an undesirable light is not categorized G unless it also falls in one of the criteria listed for G.

COVERT RESISTANCE

Covert resistance is defined as the extent to which the client manifests indirect or impersonalized criticism of the therapist or therapy, also blocking, delaying tactics, failure to recall or report things, changing the subject, interrupting the therapist. It is resistance or hostility toward therapy, therapist, progress in therapy, or toward things which are thought of as being conducive to such progress. But the resistance is not directly expressed verbally; instead, other more subtle escapes or hostilities are resorted to by the client.

1. Long Pauses (LP): A 15-second "response" in which the client does not talk.

2. Short answer (SA): An unelaborated simple thought statement not longer than four or five words that is followed by a pause. This category is not used when the client's response is in reply to a question by the therapist which can be given a simple affirmative, negative, or factual answer.

3. Changing Topic (TC): Client initiated changes in the topic being discussed. The new topic is clearly unrelated to the previous statements of the counselor or client. Client statements which serve to cut off a topic (e.g., "That's all I have to say about it.") are also included here.

4. Blocking (BI): Incomplete sentences that are not the product of an interruption by the therapist. Retracing and rephrasing or fumbling of sentences. Pauses

in mid-sentence (at least 5 seconds in length). Any statement of, or indication of, failure or inability to think or talk about something in particular or anything in general; inability to recall a word, situation, or example; inability or failure to give reasons or motivations; inability to give any label to one's own feelings, attitudes, or perceptions.

5. Interruption (Int.): Client interruptions or overriding of therapist's verbalizations. (Agreement inserted into a therapist's verbalizations with no intent to cut him off are not included here.)

6. Verbalization or intellectualization (V): Talking about minutiae or irrelevant details. *Abstract* discussions of politics, religion, etc. Such statements bear no discernible relationship to the client's problems as he has been expressing them in the interview. To qualify for a categorization here such a statement must take up a whole 15-second interval. However, if the client spends time groping for an irrelevant detail his response is categorized here whether or not the whole 15 seconds is spent on irrelevancies.

7. Resistance toward therapy or therapist (Th): Indirect or disguised reference to the inadequacy of therapy or therapist. This includes indirect or subtle minimizations of the benefits being derived from therapy. These statements are not made with any reference to the client's own feelings or thoughts about therapy or therapist, or the client is not taking any definite stand or position in his statement.

Statements indicating unwillingness to abide by therapeutic limits or demands when this unwillingness is only indirectly expressed. Any reason—other than open admission of not wanting to do so—for not appearing on time. Indecision about attending the next interview without stating any desire not to.

8. Inability to understand (U): Client statements indicating feigned or real inability to understand the therapist when the therapist's statement is judged to be clear in meaning. Requests for clarification of such therapist statements are included here, but not requests for further elaboration (which are included in Dependence).

Note: A rejection of a therapist's statement or interpretation is not considered resistance, unless it also happens to meet one of the above criteria.

The following statements are some examples to clarify the categorization of Covert Resistance:

1. "I can't define how I felt in that situation." . . . CR, "BI" (Blocking).

2. Therapist: "I wonder where they get that idea?" Client: "I don't know." . . . N. Client's inability to provide labels, motivations, etc., for other people is not considered resistance.

3. "If I could be hypnotized that would help." . . . CR. This is an indirect reference to inadequacy of therapy.

4. "And I was going to tell him off, so I went down there about two o'clock (pause), or was it three o'clock (pause), I think it was two o'clock, and I said to him." . . . CR. This is "Verbalization," grouping for irrelevant detail.

5. "Since I talked about this thing with you I felt much better in class than I had before, but I think that was because I was sitting somewhere else this time." . . . CR ("Th"). This is indirect minimization of the benefits being derived from therapy.

6. Therapist: "You felt uncomfortable in that situation." Client: "What did you say?" . . . CR. This is real or feigned inability to understand a clear therapist statement.

7. "I don't know if I'll be able to make it next time, I have an exam coming up." . . . CR ("Th"). Client gives reason, other than not wanting to, for possibility of not coming to therapy session.

OPEN RESISTANCE

Briefly defined, "open resistance" is the extent to which the client verbalizes criticism—in an open way—of the therapist or the therapeutic technique. Also, per-

sonal and verbalized opposition to staying within the limits set by the particular kind of therapy which the client is receiving. This is *verbalized unwillingness* as opposed to "inability" or failure per se.

This is resistance or hostility toward the same things as are mentioned in "covert resistance," but here these things are admitted by the client.

1. Criticism or negative attitudes about the therapist or therapy verbally expressed. Statements are frequently expressed in the form of doubts, sarcastic remarks, and only thinly veiled criticism of the therapy or therapist. These statements admittedly convey the thoughts or feelings of the client himself and the client is taking a stand about his opinions or doubts.

2. Open admission of unwillingness (not "inability") to talk, or discuss any particular area, or to follow

the conditions and limits of the kind of therapy the client is receiving.

The following statements are some examples to clarify the categorization of Open Resistance:

1. "I can't see how this is going to help very much." . . . Op. Res.
2. "All that's been happening so far is I've been answering questions, when am I going to get some answers?" . . . Op. Res.
3. "Isn't there some way we could speed this business up?" . . . Op. Res.
4. "That's something I'd rather not talk about right now" . . . Op. Res.
5. "I'd like to take a break from this for a while. I'll call you again when I want another appointment if that's O.K. with you." . . . Op. Res.

THE TEST OF CLINICAL JUDGMENT

DIRECTIONS

This is a test of your clinical judgment. It consists of statements made by all sorts of persons. Your task is to read each statement carefully and then make a judgment about the *emotional adjustment of the person who made the statement*. Do not evaluate the content of the statement itself.

If you can conceive of a statement as coming from an essentially normal person, place a check under the column marked "A." If you cannot conceive of the statement as coming from an essentially normal person, place a check under the column marked "B." Place one check in front of each statement.

"NORMAL" FILLER ITEMS^a

1. Money can become an important thing in anyone's life when he's really short of it.
2. Farmers have as much common sense as other people.
3. The threat of pneumonia is much less marked today than it once was.
4. A person doesn't like to think that his home town has turned out to be a dump.
5. One of the roughest things about the military life is that you have to do so many apparently senseless things.
6. It's hard for most of us to realize what it is like to go for days without getting enough food.

"ABNORMAL" FILLER ITEMS

1. Little by little, drop by drop, the doctors drain the blood from you 'til you are completely bloodless.
2. The men in Moscow have mind machines which they use to control anybody they want to control.
3. Everybody should really shower 10 or 12 times a day to be absolutely sure that his body is free of killer germs.
4. It's because they've been so sinful that people's bodies start to rot inside and you can smell them.
5. They should send all of the iceboxes and refrigerators to Alaska and everyone should go up there and live in one with his mother and father.
6. Everyone is doomed. Very soon now, the earth will open up and we'll all be swallowed into Hell.

TEST ITEMS

1. My life has been laid out for me and I can do no more than live it as has been planned.
2. Wives should be traded around at fairly frequent intervals. Otherwise things get boring.
3. You can usually tell whether a kid is going to turn out to be a bad egg or not and if he is you might as well lock him up right then. There wouldn't be nearly so many killings and things if we'd lock them up and get rid of them.
4. The ordinary person doesn't know what's good for

^aGrouped for the reader's convenience.

- him and what isn't. He needs to be pushed for his own good.
5. Don't try to tell me that most everybody wouldn't cheat on tests if they could get away with it. They're just too scared, that's all.
6. Rich people are generally sex perverts or free-lovers.
7. We women are useless for anything except bearing children and keeping house.
8. Good and evil are constantly battling for your soul as well as mine.
9. Old maids are frustrated old women who get so eager for sex that they practice a lot of perversions.
10. Have you ever heard of them putting a Morgan or Rockefeller in jail? You're damned right they don't. Only the little guys pay.
11. Survival of the fittest is the law of God. It is wrong for us to try to keep weaklings alive.
12. Only a woman can really understand how others feel about love. Men are such insensitive brutes.
13. If a man can't hold down an honest job, then I say let him starve.
14. I advocate celibacy. I feel that there's little place for sex in professional life.
15. It's the devil that makes people do bad things. He's the one who is behind all of the bad people.
16. Women who paint up their eyes and faces must be whores at heart.
17. All of this nonsense about not frustrating children will lead to no good. God meant for us to have to do things we don't want to do or he wouldn't have set things up the way he did.
18. No rich man could get rich without stepping all over a lot of people on the way.
19. If I know a person is dishonest, even once, I'll never trust him again.
20. Yeah, they make more money at their fancy desks than you do out here in the yard, but did you ever look close at them? They're soft and white all over—just like women. They're nothing but a bunch of fruits.
21. Getting the jump on the next guy so that you end up on top is the main thing about living in this world, the way it is.
22. It's dangerous to let young boys and girls get together by themselves. The first thing they'll do is start playing around with each other.
23. People who get sick should pray to God to forgive them for their sins. Sickness is God's way of punishing us for our wicked ways.
24. It's gotta be a pretty queer guy in my way of thinking who would ever knit anything.
25. They just fill kids' minds with a lot of fancy stuff at those colleges so that their old men aren't good enough for them anymore.
26. You get a wild kid—the only way to handle him is to break him. Otherwise he'll end up a crook or something.

27. Listen, take my advice. Don't be fooled by the soft, delicate appearance women make. They'd take over this world and make slaves of us men if we'd relax our control just one little bit.
28. They ought to open the floodgates and flood the whole damned South with everyone in it.
29. Don't talk to me about those clergymen, with their "holier than thou" attitudes. They'd lie, cheat, and steal just as much as anybody.
30. Parents always know what's right, so it's best not to do anything without first asking them.
31. Most college graduates are radicals, homosexuals, or troublemakers.
32. Girls who go around screwing with the boys ought to be whipped in public and made an example of.
33. Those people in the slums are basically weak in character and you'll never make them strong like good honest workers by giving them more parks and fresh air.
34. Every big corporation and most of the smaller ones have company spies mixed in with the workers.
35. You've really got to make a criminal suffer severely for his crime. Punish him until he has no spirit left. Then he'll be willing to do what is right.
36. Dancing and card playing are just as sinful as sex. Better a person should keep himself locked in his room than to indulge in these evils.
37. My mother always said to act like a lady at all times. She is right. I've found that if a woman lets her guard down for a minute that a "gentleman" will change to an animal, just like that.
38. Man is basically evil and at the most can do no more than to control the evil that is within him.
39. It's bad business letting women take over men's jobs. Why do you think that they hate to have them in mines and on ships? Do you think it's just superstition?
40. A child must learn who his master is. There's plenty of time for him to think and act for himself when he's an adult.
41. If you don't stop those people from talking with each other, you will soon find that they have formed some sort of conspiracy which you will have a hard time suppressing.
42. We should destroy criminals because they'll just produce more of their own kind.
43. I tell you it's bad to send 18-year-olds to the Army. They'll turn them out as killers, cynics, homosexuals—or worse.
44. Laborers are a bunch of greasy, dirty slobs.
45. If it was a matter of sacrificing some of my happiness to help out someone else, I wouldn't do it.
46. Who would criticize the President of the United States except Communists, homosexuals, extremists, and that sort.
47. Who do these women think they're kidding with all of their modesty and morals? They'd jump into bed with the first guy that asked them if they thought they could get away with it.
48. Killing dogs and monkeys is a horrible thing—no matter how much good they claim will come of it.
49. The big shots can get away with anything. But just don't let the little man step over the line.
50. Bachelors are like vicious animals waiting to pounce on any young girl they can and take advantage of her.
51. I don't see why we should tax the people who are smart enough to make out in order to take care of those who are too dumb or too damned lazy to take care of themselves.
52. When somebody does not enforce a rule, it adds one more bit of disorganization to things. It's things like that which create no end of trouble for everybody.
53. You gotta be a bigger crook than they are; that's the only way you'll get the respect of those big shots.
54. There is no such thing as a "white lie." Any deviation from the exact truth, no matter how insignificant it may seem, fosters evil.
55. Don't let 'em fool you with these "I want to help you" spiels. They got an angle and they just want to use you.
56. A woman who would work as a bartender must be nothing but a prostitute. What other reason would she have for taking a job like that?
57. A company union, ha! You know what that's for don't you? So that the bosses can run it and bleed you for all you're worth.

POSTTHERAPY RATING SCALE

DIRECTIONS

In rating the client on each item of the scale, consider the difference between the way the client was when he entered therapy and the way he was at the end of treatment. Base your rating on the change that has occurred in the client within this interval.

On certain items you may feel that a client came into therapy with a behavior already incorporated as part of his usual behavior and that he has not changed in this particular area. On these items, clients will be rated as "remained the same." For example, a client may have been quite willing to discuss the significant aspects of his problems from the time he entered therapy, and he has made no change in this respect.

The scale consists of two kinds of items. Most of the items deal with behavior which is not specifically related to the counseling situation; however, certain items are based on the client's behavior in therapy.

On those items which are related to over-all or general behavior, your rating should reflect changes in the behavior and/or characteristics of the client which you feel are relatively permanent and which also apply to his behavior in situations outside of therapy. On these items indicate *only* changes which you feel the client has integrated into his general behavior.

On those items related to behavior in therapy, try to distinguish mere conformity to the demands of therapy from those changes which are real reorganizations of the

client's attitudes, feelings, and motivations operating within therapy. Your rating should *not* be based on mere conformity.

In certain cases, you are asked to rate changes in the client's behavior in terms of whether they are realistic or appropriate. For example, if a client was too positive in his evaluation of other people, a more realistic position might have resulted from the client's developing a more critical attitude toward other people. If the client had been hypercritical of others, a more realistic position would be based on the extent to which the client may have become less critical of others.

ITEMS

1. The client's perception of the problem as a function of his own behavior has:
 - (a) become less realistic; (b) remained the same; (c) become slightly more realistic; (d) become moderately more realistic; (e) become considerably more realistic.
2. The client's placing of responsibility for his difficulties on others and/or environmental circumstances has:
 - (a) become considerably more realistic; (b) become moderately more realistic; (c) become slightly more realistic; (d) remained the same; (e) become less realistic.
3. The extent to which the client perceives the prob-

- lem merely in symptomatic terms has:
 (a) decreased considerably; (b) decreased moderately; (c) decreased slightly; (d) remained the same; (e) increased.
4. The client's feelings of discomfort in his everyday life have: (a) decreased considerably; (b) decreased moderately; (c) decreased slightly; (d) remained the same; (e) increased.
 5. The client's expression of positive emotions when they are appropriate has:
 (a) decreased; (b) remained the same; (c) increased slightly; (d) increased moderately; (e) increased considerably.
 6. The client's symptoms have:
 (a) become considerably less disturbing; (b) become moderately less disturbing; (c) become slightly less disturbing; (d) remained the same; (e) become more disturbing.
 7. The client discusses feelings and attitudes which are relevant to the problem:
 (a) considerably more freely; (b) moderately more freely; (c) slightly more freely; (d) the same; (e) less freely.
 8. The client's comprehension of important causal relationships between his symptoms and the underlying needs and conflicts has:
 (a) decreased; (b) remained the same; (c) increased slightly; (d) increased moderately; (e) increased considerably.
 9. The client's avoidance of making decisions which would seem to be necessary before anything can be accomplished has:
 (a) decreased considerably; (b) decreased moderately; (c) decreased slightly; (d) remained the same; (e) increased.
 10. The client's attitudes toward other people have:
 (a) become less appropriate; (b) remained the same; (c) become slightly more appropriate; (d) become moderately more appropriate; (e) become considerably more appropriate.
 11. Concerning his ability to solve his problems, the client has:
 (a) become less confident; (b) become no more confident; (c) become slightly more confident; (d) become moderately more confident; (e) become considerably more confident.
 12. The client's understanding of his problems in terms of his own past experience has:
 (a) become considerably more meaningful; (b) become moderately more meaningful; (c) become slightly more meaningful; (d) remained the same; (e) become less meaningful.
 13. The client feels his relationships with others have:
 (a) become less satisfactory; (b) remained the same; (c) become slightly more satisfactory; (d) become moderately more satisfactory; (e) become considerably more satisfactory.
 14. The extent to which the client develops new plans for improving his situation has:
 (a) increased considerably; (b) increased moderately; (c) increased slightly; (d) remained the same; (e) decreased.
 15. The client's feelings and attitudes toward himself have:
 (a) become considerably more appropriate; (b) become moderately more appropriate; (c) become slightly more appropriate; (d) remained the same; (e) become less appropriate.
 16. The client's expression of negative emotions when they are appropriate has: (a) decreased; (b) remained the same; (c) increased slightly; (d) increased moderately; (e) increased considerably.
 17. The client's view of his strengths and shortcomings has:
 (a) become less realistic; (b) remained the same; (c) become slightly more realistic; (d) become moderately more realistic; (e) become considerably more realistic.
 18. The client's emotional reactions to other people and situations have:
 (a) become considerably more realistic; (b) become moderately more realistic; (c) become slightly more realistic; (d) remained the same; (e) become less realistic.
 19. The client's understanding of his problems in terms of his own needs has:
 (a) become considerably more accurate; (b) become moderately more accurate; (c) become slightly more accurate; (d) remained the same; (e) become less accurate.
 20. The client's expectations for himself in the interpersonal aspects of his life have:
 (a) become less realistic; (b) remained the same; (c) become slightly more realistic; (d) become moderately more realistic; (e) become considerably more realistic.
 21. The client's acceptance of responsibility for solving his problems has:
 (a) increased considerably; (b) increased moderately; (c) increased slightly; (d) remained the same; (e) decreased.
 22. The client's perception of the meaning of the reactions of others toward him has:
 (a) become considerably less distorted; (b) become moderately less distorted; (c) become slightly less distorted; (d) remained the same; (e) become more distorted.
 23. The client's meaningful emotional involvement as opposed to an intellectual approach to his problems has:
 (a) increased considerably; (b) increased moderately; (c) increased slightly; (d) remained the same; (e) decreased.
 24. The client's attempts to avoid significant areas of discussion has:
 (a) increased; (b) remained the same; (c) decreased slightly; (d) decreased moderately; (e) decreased considerably.
 25. The extent to which the client sees where his own characteristic ways of thinking, feeling, and behaving bear an important relationship to his problems has:
 (a) decreased; (b) remained the same; (c) increased slightly; (d) increased moderately; (e) increased considerably.
 26. The client's understanding of his problems as resulting from interpersonal relationships with other people has:
 (a) become less realistic; (b) remained the same; (c) become slightly more realistic; (d) become moderately more realistic; (e) become considerably more realistic.
 27. The client's attempts to try new ways of handling his problems have:
 (a) decreased; (b) remained the same; (c) increased slightly; (d) increased moderately; (e) increased considerably.



